RESEARCH CENTER

FIELD

Activity Report 2014

Section Scientific Foundations

Edition: 2015-03-24

ALGORITHMICS, PROGRAMMING, SOFTWARE AND ARCHITECTURE

ALF Project-Team

3. Research Program

3.1. Motivations

Multicores have become mainstream in general-purpose as well as embedded computing in the last few years. The integration technology trend allows to anticipate that a 1000-core chip will become feasible before 2020. On the other hand, while traditional parallel application domains, e.g. supercomputing and transaction servers, are benefiting from the introduction of multicores, there are very few new parallel applications that have emerged during the last few years.

In order to allow the end-user to benefit from the technological breakthrough, new architectures have to be defined for the 2020's many-cores, new compiler and code generation techniques as well as new performance prediction/guarantee techniques have to be proposed .

3.2. The context

3.2.1. Technological context: The advent of multi- and many- core architecture

For almost 30 years since the introduction of the first microprocessor, the processor industry was driven by the Moore's law till 2002, delivering performance that doubled every 18-24 months on a uniprocessor. However since 2002 , and despite new progress in integration technology, the efforts to design very aggressive and very complex wide issue superscalar processors have essentially been stopped due to poor performance returns, as well as power consumption and temperature walls.

Since 2002-2003, the microprocessor industry has followed a new path for performance: the so-called multicore approach, i.e., integrating several processors on a single chip. This direction has been followed by the whole processor industry. At the same time, most of the computer architecture research community has taken the same path, focusing on issues such as scalability in multicores, power consumption, temperature management and new execution models, e.g. hardware transactional memory.

In terms of integration technology, the current trend will allow to continue to integrate more and more processors on a single die. Doubling the number of cores every two years will soon lead to up to a thousand processor cores on a single chip. The computer architecture community has coined these future processor chips as many-cores.

3.2.2. The application context: multicores, but few parallel applications

For the past five years, small scale parallel processor chips (hyperthreading, dual and quad-core) have become mainstream in general-purpose systems. They are also entering the high-end embedded system market. At the same time, very few (scalable) mainstream parallel applications have been developed. Such development of scalable parallel applications is still limited to niche market segments (scientific applications, transaction servers).

3.2.3. The overall picture

Till now, the end-user of multicores is experiencing improved usage comfort because he/she is able to run several applications at the same time. Eventually, in the near future with the 8-core or the 16-core generation, the end-user will realize that he/she is not experiencing any functionality improvement or performance improvement on current applications. The end-user will then realize that he/she needs more effective performance rather than more cores. The end-user will then ask either for parallel applications or for more effective performance on sequential applications.

3.3. Technology induced challenges

3.3.1. The power and temperatures walls

The power and the temperature walls largely contributed to the emergence of the small-scale multicores. For the past five years, mainstream general-purpose multicores have been built by assembling identical superscalar cores on a chip (e.g. IBM Power series). No new complex power hungry mechanisms were introduced in the core architectures, while power saving techniques such as power gating, dynamic voltage and frequency scaling were introduced. Therefore, since 2002, the designers have been able to keep the power consumption budget and the temperature of the chip within reasonable envelopes while scaling the number of cores with the technology.

Unfortunately, simple and efficient power saving techniques have already caught most of the low hanging fruits on energy consumption. Complex power and thermal management mechanisms are now becoming mainstream; e.g. the Intel Montecito (IA64) featured an adjunct (simple) core whose unique mission is to manage the power and temperature on two cores. Processor industry will require more and more heroic efforts on this power and temperature management policy to maintain its current performance scaling path. Hence the power and temperature walls might slow the race towards 100's and 1000's cores unless the processor industry takes a new paradigm shift from the current "replicating complex cores" (e.g. Intel Nehalem) towards many simple cores (e.g. Intel Larrabee) or heterogeneous manycores (e.g. new GPUs, IBM Cell).

3.3.2. The memory wall

For the past 20 years, the memory access time has been one of the main bottlenecks for performance in computer systems. This was already true for uniprocessors. Complex memory hierarchies have been defined and implemented in order to limit the visible memory access time as well as the memory traffic demands. Up to three cache levels are implemented for uniprocessors. For multi- and many-cores the problems are even worse. The memory hierarchy must be replicated for each core, memory bandwidth must be shared among the distinct cores, data coherency must be maintained. Maintaining cache coherency for up to 8 cores can be handled through relatively simple bus protocols. Unfortunately, these protocols do not scale for large numbers of cores, and there is no consensus on coherency mechanism for manycore systems. Moreover there is no consensus on core organization (flat ring? flat grid? hierarchical ring or grid?).

Therefore, organizing and dimensioning the memory hierarchy will be a major challenge for the computer architects. The successful architecture will also be determined by the abilitty of the applications (i.e., the programmers or the compilers or the run-time) to efficiently place data in the memory hierarchy and achieve high performance.

Finally new technology opportunities may demand to revisit the memory hierarchy. As an example, 3D memory stacking enables a huge last-level cache (maybe several gigabytes) with huge bandwidth (several Kbits/ processor cycle). This dwarfs the main memory bandwidth and may lead to other architectural tradeoffs.

3.4. Need for efficient execution of parallel applications

Achieving high performance on future multicores will require the development of parallel applications, but also an efficient compiler/runtime tool chain to adapt codes to the execution platform.

3.4.1. The diversity of parallelisms

Many potential execution parallelism patterns may coexist in an application. For instance, one can express some parallelism with different tasks achieving different functionalities. Within a task, one can expose different granularities of parallelism; for instance a first layer message passing parallelism (processes executing the same functionality on different parts of the data set), then a shared memory thread level parallelism and fine grain loop parallelism (a.k.a vector parallelism).

Current multicores already feature hardware mechanisms to address these different parallelisms: physically distributed memory — e.g. the new Intel Nehalem already features 6 different memory channels — to address task parallelism, thread level parallelism — e.g. on conventional multicores, but also on GPUs or on Cell-based machines —, vector/SIMD parallelism — e.g. multimedia instructions. Moreover they also attack finer instruction level parallelism and memory latency issues. Compilers have to efficiently discover and manage all these forms to achieve effective performance.

3.4.2. Portability is the new challenge

Up to now, most parallel applications were developed for specific application domains in high end computing. They were used on a limited set of very expensive hardware platforms by a limited number of expert users. Moreover, they were executed in batch mode.

In contrast, the expectation of most end-users of the future mainstream parallel applications running on multicores will be very different. The mainstream applications will be used by thousands, maybe millions of non-expert users. These users consider functional portability of codes as granted. They will expect their codes to run faster on new platforms featuring more cores. They will not be able to tune the application environment to optimize performance. Finally, multiple parallel applications may have to be executed concurrently.

The variety of possible hardware platforms, the lack of expertise of the end-users and the varying run-time execution environments will represent major difficulties for applications in the multicore era.

First of all, the end user considers functional portability without recompilation as granted, this is a major challenge on parallel machines. Performance portability/scaling is even more challenging. It will become inconceivable to rewrite/retune each application for each new parallel hardware platform generation to exploit them. Therefore, apart from the initial development of parallel applications, the major challenge for the next decade will be to *efficiently* run parallel applications on hardware architectures radically different from their original hardware target.

3.4.3. The need for performance on sequential code sections

3.4.3.1. Most software will exhibit substantial sequential code sections

For the foreseeable future, the majority of applications will feature important sequential code sections.

First, many legacy codes were developed for uniprocessors. Most of these codes will not be completely redeveloped as parallel applications, but will evolve to applications using parallel sections for the most compute-intensive parts. Second, the overwhelming majority of the programmers have been educated to program in a sequential programming style. Parallel programming is much more difficult, time consuming and error prone than sequential programming. Debugging and maintaining a parallel code is a major issue. Investing in the development of a parallel application will not be cost-effective for the vast majority of software developments. Therefore, sequential programming style will continue to be dominant in the foreseeable future. Most developers will rely on the compiler to parallelize their application and/or use some software components from parallel libraries.

3.4.3.2. Future parallel applications will require high performance sequential processing on 1000's cores chip

With the advent of universal parallel hardware in multicores, large diffusion parallel applications will have to run on a broad spectrum of parallel hardware platforms. They will be used by non-expert users who will not be able to tune the application environment to optimize performance. They will be executed concurrently with other processes which may be interactive.

The variety of possible hardware platforms, the lack of expertise of the end-user and the varying run-time execution environments are major difficulties for parallel applications. This tends to constrain the programming style and therefore reinforces the sequential structure of the control of the application.

Therefore, *most future parallel applications will rely on a single main thread or a few main threads in charge of distinct functionalities of the application. Each main thread will have a general sequential control and can initiate and control the parallel execution of parallel tasks.*

In 1967, Amdahl [37] pointed out that, if only a portion of an application is accelerated, the execution time cannot be reduced below the execution time of the residual part of the application. Unfortunately, even highly parallelized applications exhibit some residual sequential part. For parallel applications, this indicates that the effective performance of the future 1000's cores chip will significantly depend on their ability to be efficient on the execution of the control portions of the main thread as well as on the execution of sequential portions of the application.

3.4.3.3. The success of 1000's cores architecture will depend on single thread performance

While the current emphasis of computer architecture research is on the definition of scalable multi- many- core architectures for highly parallel applications, we believe that the success of the future 1000-core architecture will depend not only on their performance on parallel applications including sequential sections, but also on their performance on single thread workloads.

3.5. Performance evaluation/guarantee

Predicting/evaluating the performance of an application on a system without explicitly executing the application on the system is required for several usages. Two of these usages are central to the research of the ALF project-team: microarchitecture research (the system to be be evaluated does not exist) and Worst Case Execution Time estimation for real-time systems (the numbers of initial states or possible data inputs is too large).

When proposing a micro-architecture mechanism, its impact on the overall processor architecture has to be evaluated in order to assess its potential performance advantages. For microarchitecture research, this evaluation is generally done through the use of cycle-accurate simulation. Developing such simulators is quite complex and microarchitecture research was helped but also biased by some popular public domain research simulators (e.g. Simplescalar [38]). Such simulations are CPU consuming and simulations cannot be run on a complete application. Sampling representative slices of the application was proposed [4] and popularized by the Simpoint [48] framework.

Real-time systems need a different use of performance prediction; on hard real-time systems, timing constraints must be respected independently from the data inputs and from the initial execution conditions. For such a usage, the Worst Case Execution Time (WCET) of an application must be evaluated and then checked against the timing constraints. While safe and tight WCET estimation techniques and tools exist for reasonably simple embedded processors (e.g. techniques based on abstract interpretation such as [40]), accurate evaluation of the WCET of an algorithm on a complex uniprocessor system is a difficult problem. Accurately modelling data cache behavior [3] and complex superscalar pipelines are still research questions as illustrated by the presence of so-called *timing anomalies* in dynamically scheduled processors, resulting from complex interactions between processor elements (among others, interactions between caching and instruction scheduling) [45].

With the advance of multicores, evaluating / guaranteeing a computer system response time is becoming much more difficult. Interactions between processes occurs at different levels. The execution time on each core depends on the behavior of the other cores. Simulations of 1000's cores micro-architecture will be needed in order to evaluate future many-core proposals. While a few multiprocessor simulators are available for the community, these simulators cannot handle realistic 1000's cores micro-architecture. New techniques have to be invented to achieve such simulations. WCET estimations on multicore platforms will also necessitate radically new techniques, in particular, there are predictability issues on a multicore where many resources are shared; those resources include the memory hierarchy, but also the processor execution units and all the hardware resources if SMT is implemented [52].

3.6. General research directions

The overall performance of a 1000's core system will depend on many parameters including architecture, operating system, runtime environment, compiler technology and application development. In the ALF project, we will essentially focus on architecture, compiler/execution environment as well as performance

predictability, and in particular WCET estimation. Moreover, architecture research, and to a smaller extent, compiler and WCET estimation researches rely on processor simulation. A significant part of the effort in ALF will be devoted to define new processor simulation techniques.

3.6.1. Microarchitecture research directions

We have identified that high performance on single threads and sequential codes is one of the key issues for enabling overall high performance on a 1000's core system and we anticipate that the general architecture of such 1000's core chip will feature many simple cores and a few very complex cores.

Therefore our research in the ALF project will focus on refining the microarchitecture to achieve high performance on single process and/or sequential code sections within the general framework of such an heteregeneous architecture. This leads to two main research directions 1) enhancing the microarchitecture of high-end superscalar processors, 2) exploiting/modifying heterogeneous multicore architecture on a single process. The temperature wall is also a major technological/architectural issue for the design of future processor chips.

3.6.1.1. Enhancing complex core microarchitecture

Research on wide issue superscalar processors was merely stopped around 2002 due to limited performance returns and the power consumption wall.

When considering a heterogeneous architecture featuring hundreds of simple cores and a few complex cores, these two obstacles will partially vanish: 1) the complex cores will represent only a fraction of the chip and a fraction of its power consumption. 2) any performance gain on (critical) sequential threads will result in a performance gain of the whole system

On the complex core, the performance of a sequential code is limited by several factors. At first, on current architectures, it is limited by the peak performance of the processor. To push back this first limitation, we will explore new microarchitecture mechanisms to increase the potential peak performance of a complex core enabling larger instruction issue width. The processor performance is also limited by control dependencies. To push back this limitation, we will explore new branch prediction mechanisms as well as new directions for reducing branch misprediction penalties [10], [12]. As data dependencies may strongly limit performance, we will revisit data prediction. Processor performance is also often highly dependent on the presence or absence of data in a particular level of the memory hierarchy. For the ALF multicore, we will focus on sharing the access to the memory hierarchy in order to adapt the performance of the main thread to the performance of the other cores. All these topics should be studied with the new perspective of quasi unlimited silicon budget.

3.6.1.2. Exploiting heterogeneous multicores on single process

When executing a sequential section on the complex core, the simple cores will be free. Two main research directions to exploit thread level parallelism on a sequential thread have been initiated in late 90's within the context of simultaneous multithreading and early chip multiprocessor proposals: helper threads and speculative multithreading.

Helper threads were initially proposed to improve the performance of the main threads on simultaneous multithreaded architectures [39]. The main idea of helper threads is to execute codes that will accelerate the main thread without modifying its semantic.

In many cases, the compiler cannot determine if two code sections are independent due to some unresolved memory dependency. When no dependency occurs at execution time, the code sections can be executed in parallel. Thread-Level Speculation has been proposed to exploit coarse grain speculative parallelism. Several hardware-only proposals were presented [46], but the most promising solutions integrate hardware support for software thread-level speculation [50].

In the context of future manycores, thread-level speculation and helper threads should be revisited. Many simple cores will be available for executing helper threads or speculative thread execution during the execution of sequential programs or sequential code sections. The availability of these many cores is an opportunity as well as a challenge. For example, one can try to use the simple cores to execute many different helper threads

that could not be implemented within a simultaneous multithreaded processor. For thread level speculation, the new challenge is the use of less powerful cores for speculative threads. Moreover the availability of many simple cores may lead to the use of helper threads and thread level speculation at the same time.

3.6.1.3. Temperature issues

Temperature is one of the constraints that have prevented the processor clock frequency to be increased in recent years. Besides techniques to decrease the power consumption, the temperature issue can be tackled with *dynamic thermal management* [9] through techniques such as clock gating or throttling and *activity migration* [49][5].

Dynamic thermal management (DTM) is now implemented on existing processors. For high performance, processors are dimensioned according to the average situation rather than to the worst case situation. Temperature sensors are used on the chip to trigger dynamic thermal management actions, for instance thermal throttling whenever necessary. On multicores, it is possible to migrate the activity from one core to another in order to limit temperature.

A possible way to increase sequential performance is to take advantage of the smaller gate delay that comes with miniaturization, which permits in theory to increase the clock frequency. However increasing the clock frequency generally requires to increase the instantaneous power density. This is why DTM and activity migration will be key techniques to deal with Amdahl's law in future many-core processors.

3.6.2. Processor simulation research

Architecture studies, and in particular microarchitecture studies, require extensive validations through detailed simulations. Cycle accurate simulators are needed to validate the microarchitectural mechanisms.

Within the ALF project, we can distinguish two major requirements on the simulation: 1) single process and sequential code simulations 2) parallel code sections simulations.

For simulating parallel code sections, a cycle-accurate microarchitecture simulator of a 1000-core architecture will be unacceptably slow. In [6], we showed that mixing analytical modeling of the global behavior of a processor with detailed simulation of a microarchitecture mechanism allows to evaluate this mechanism. Karkhanis and Smith [42] further developed a detailed analytical simulation model of a superscalar processor. Building on top of these preliminary researches, simulation methodology mixing analytical modeling of the simple cores with a more detailed simulation of the complex cores is appealing. The analytical model of the simple cores will aim at approximately modeling the impact of the simple core execution on the shared resources (e.g. data bandwidth, memory hierarchy) that are also used by the complex cores.

Other techniques such as regression modeling [43] can also be used for decreasing the time required to explore the large space of microarchitecture parameter values. We will explore these techniques in the context of many-core simulation.

In particular, research on temperature issues will require the definition and development of new simulation tools able to simulate several minutes or even hours of processor execution, which is necessary for modeling thermal effects faithfully.

3.6.3. Compiler research directions

3.6.3.1. General directions

Compilers are keystone solutions for any approach that deals with high performance on 100+ processors systems. But general-purpose compilers try to embrace so many domains and try to serve so many constraints that they frequently fail to achieve very high performance. They need to be deeply revisited. We identify four main compiler/software related issues that must be addressed in order to allow efficient use of multi- and many-cores: 1) programming 2) resource management 3) application deployment 4) portable performance. Addressing these challenges will require to revisit parallel programming and code generation extensively.

The past of parallel programming is scattered with hundreds of parallel languages. Most of these languages were designed to program homogeneous architectures and were targeting a small and well-trained community of HPC programmers. With the new diversity of parallel hardware platforms and the new community of non-expert developers, expressing parallelism is not sufficient anymore. Resource management, application deployment and portable performance are intermingled issues that require to be addressed holistically.

As many decisions should be taken according to the available hardware, resource management cannot be separated from parallel programming. Deploying applications on various systems without having to deal with thousands of hardware configurations (different numbers of cores, accelerators, ...) will become a major concern for software distribution. The grail of parallel computing is to be able to provide portable performance on a large set of parallel machines and varying execution contexts.

Recent techniques are showing promises. Iterative compilation techniques, exploiting the huge CPU cycle count now available, can be used to explore the optimization space at compile-time. Second, machine-learning techniques can be used to automatically improve compilers and code generation strategies. Speculation can be used to deal with necessary but missing information at compile-time. Finally, dynamic techniques can select or generate at run-time the most efficient code adapted to the execution context and available hardware resources.

Future compilers will benefit from past research, but they will also need to combine static and dynamic techniques. Moreover, domain specific approaches might be needed to ensure success. The ALF research effort will focus on these static and dynamic techniques to address the multicore application development challenges.

3.6.3.2. Portability of applications and performance through virtualization

The life cycle is much longer for applications than for hardware. Unfortunately the multicore era jeopardizes the old binary compatibility recipe. Binaries cannot automatically exploit additional computing cores or new accelerators available on the silicon. Moreover maintaining backward binary compatibility on future parallel architectures will rapidly become a nightmare, applications will not run at all unless some kind of dynamic binary translation is at work.

Processor virtualization addresses the problem of portability of functionalities. Applications are not compiled to the final native code but to a target independent format. This is the purpose of languages such as Java and .NET. Bytecode formats are often *a priori* perceived as inappropriate for performance intensive applications and for embedded systems. However, it was shown that compiling a C or C++ program to a bytecode format produces a code size similar to dense instruction sets [2]. Moreover, this bytecode representation can be compiled to native code with performance similar to static compilation [1]. Therefore processor virtualization for high performance, i.e., for languages like C or C++, provides significant advantages: 1) it simplifies software engineering with fewer tools to maintain and upgrade; 2) it allows better code readability and easier code maintenance since it avoids code specialization for specific targets using compile time macros such as #ifdef ; 3) the *execution code* deployed on the system is the execution code that has been debugged and validated, as opposed to the same *source code* has been recompiled for another platform; 4) new architectures will come with their JIT compiler. The JIT will (should) automatically take advantage of new architecture features such as SIMD/vector instructions or extra processors.

Our objective is to enrich processor virtualization to allow both functional portability and high performance using JIT at runtime, or bytecode-to-native code offline compiler. Split compilation can be used to annotate the bytecode with relevant information that can be helpful to the JIT at runtime or to the bytecode to native code offline compiler. Because the first compilation pass occurs offline, aggressive analyses can be run and their outcomes encoded in the bytecode. For example, such information include vectorizability, memory references (in)dependencies, suggestions derived from iterative compilation, polyhedral analysis, or integer linear programming. Virtualization allows to postpone some optimizations to run time, either because they increase the code size and would increase the cost of an embedded system or because the actual hardware platform characteristics are unknown.

3.6.4. Performance predictability for real-time systems

While compiler and architecture research efforts often focus on maximizing average case performance, applications with real-time constraints do not need only high performance but also performance guarantees in all situations, including the worst-case situation. Worst-Case Execution Time estimates (WCET) need to be upper bounds of any possible execution time. The safety level required depends on the criticality of applications: missing a frame on a video in the airplane for passenger in seat 20B is less critical than a safety critical decision in the control of the airplane.

Within the ALF project, our objective is to study performance guarantees for both (i) sequential codes running on complex cores ; (ii) parallel codes running on the multicores. This results in two quite distinct problems.

For sequential code executing on a single core, one can expect that, in order to provide real-time possibility, the architecture will feature an execution mode where a given processor will be guaranteed to access a fixed portion of the shared resources (caches, memory bandwidth). Moreover, this guaranteed share could be optimized at compile time to enforce the respect of the time constraints. However, estimating the WCET of an application on a complex micro-architecture is still a research challenge. This is due to the complex interaction of micro-architectural elements (superscalar pipelines, caches, branch prediction, out-of-order execution) [45]. We will continue to explore pure analytical and static methods. However when accurate static hardware modeling methods cannot handle the hardware complexity, new probabilistic methods [44] might be needed to explore to obtain as safe as possible WCET estimates.

Providing performance guarantees for parallel applications executed on a multicore is a new and challenging issue. Entirely new WCET estimation methods have to be defined for these architectures to cope with dynamic resource sharing between cores, in particular on-chip memory (either local memory or caches) are shared, but also buses, network-on-chip and the access to the main memory. Current pure analytical methods are too pessimistic at capturing interferences between cores [53], therefore hardware-based or compiler methods such as [51] have to be defined to provide some degree of isolation between cores. Finally, similarly to simulation methods, new techniques to reduce the complexity of WCET estimation will be explored to cope with manycore architectures.

ANTIQUE Team

3. Research Program

3.1. Semantics

Semantics plays a central role in verification since it always serves as a basis to express the properties of interest, that need to be verified, but also additional properties, required to prove the properties of interest, or which may make the design of static analysis easier.

For instance, if we aim for a static analysis that should prove the absence of runtime error in some class of programs, the concrete semantics should define properly what error states and non error states are, and how program executions step from a state to the next one. In the case of a language like C, this includes the behavior of floating point operations as defined in the IEEE 754 standard. When considering parallel programs, this includes a model of the scheduler, and a formalization of the memory model.

In addition to the properties that are required to express the proof of the property of interest, it may also be desirable that semantics describe program behaviors in a finer manner, so as to make static analyses easier to design. For instance, it is well known that, when a state property (such as the absence of runtime error) is valid, it can be established using only a state invariant (i.e., an invariant that ignores the order in which states are visited during program executions). Yet searching for trace invariants (i.e., that take into account some properties of program execution history) may make the static analysis significantly easier, as it will allow it to make finer case splits, directed by the history of program executions. To allow for such powerful static analyses, we often resort to a *non standard semantics*, which incorporates properties that would normally be left out of the concrete semantics.

3.2. Abstract interpretation and static analysis

Once a reference semantics has been fixed and a property of interest has been formalized, the definition of a static analysis requires the choice of an *abstraction*. The abstraction ties a set of *abstract predicates* to the concrete ones, which they denote. This relation is often expressed with a *concretization function* that maps each abstract element to the concrete property it stands for. Obviously, a well chosen abstraction should allow expressing the property of interest, as well as all the intermediate properties that are required in order to prove it (otherwise, the analysis would have no chance to achieve a successful verification). It should also lend itself to an efficient implementation, with efficient data-structures and algorithms for the representation and the manipulation of abstract predicates. A great number of abstractions have been proposed for all kinds of concrete data types, yet the search for new abstractions is a very important topic in static analysis, so as to target novel kinds of properties, to design more efficient or more precise static analyses.

Once an abstraction is chosen, a set of *sound abstract transformers* can be derived from the concrete semantics and that account for individual program steps, in the abstract level and without forgetting any concrete behavior. A static analysis follows as a result of this step by step approximation of the concrete semantics, when the abstract transformers are all computable. This process defines an *abstract interpretation* [40]. The case of loops requires a bit more work as the concrete semantics typically relies on a fixpoint that may not be computable in finitely many iterations. To achieve a terminating analysis we then use *widening operators* [40], which over-approximates the concrete union and ensure termination.

A static analysis defined that way always terminates and produces sound over-approximations of the programs behaviors. Yet, these results may not be precise enough for verification. This is where the art of static analysis design comes into play through, among others:

- the use of more precise, yet still efficient enough abstract domains;
- the combination of application specific abstract domains;
- the careful choice of abstract transformers and widening operators.

3.3. Applications of the notion of abstraction in semantics

In the previous subsections, we sketched the steps in the design of a static analyzer to infer some family of properties, which should be implementable, and efficient enough to succeed in verifying non trivial systems.

Yet, the same principles can also be applied successfully to other goals. In particular, the abstract interpretation framework should be viewed a very general tool to *compare different semantics*, not necessarily with the goal of deriving a static analyzer. Such comparisons may be used in order to prove two semantics equivalent (i.e., one is an abstraction of the other and vice versa), or that a first semantics is strictly more expressive than another one (i.e., the latter can be viewed an abstraction of the former, where the abstraction actually makes some information redundant, which cannot be recovered). A classical example of such comparison is the classification of semantics of transition systems [38], which provides a better understanding of program semantics in general. For instance, this approach can be applied to get a better understanding of the semantics of a programming language, but also to select which concrete semantics should be used as a foundation for a static analysis, or to prove the correctness of a program transformation, compilation or optimization.

3.4. The analysis of biological models

One of our application domains, the analysis of biological models, is not a classical target of static analysis because it aims at analyzing models instead of programs. Yet, the analysis of biological models is closely intertwined with the other application fields of our group. Firstly, abstract interpretation provides a formal understanding of the abstraction process which is inherent to the modeling process. Abstract interpretation is also used to better understand the systematic approaches which are used in the systems biology field to capture the properties of models, until getting formal, fully automatic, and scalable methods. Secondly, abstract interpretation is used to offer various semantics with different grains of abstraction, and, thus, new methods to apprehend the overall behavior of the models. Conversely, some of the methods and abstractions which are developed for biological models are inspired by the analysis of concurrent systems and by security analysis. Lastly, the analysis of biological models raises issues about differential systems, stochastic systems, and hybrid systems. Any breakthrough in these directions will likely be very important to address the important challenge of the certification of critical systems in interaction with their physical environment.

AOSTE Project-Team

3. Research Program

3.1. Models of Computation and Communication (MoCCs)

Participants: Julien Deantoni, Robert de Simone, Frédéric Mallet, Jean-Vivien Millo, Dumitru Potop Butucaru.

Esterel, SyncCharts, synchronous formalisms, Process Networks, Marked Graphs, Kahn networks, compilation, synthesis, formal verification, optimization, allocation, refinement, scheduling

Formal Models of Computation form the basis of our approach to Embedded System Design. Because of the growing importance of communication handling, it is now associated with the name, MoCC in short. The appeal of MoCCs comes from the fact that they combine features of mathematical models (formal analysis, transformation, and verification) with these of executable specifications (close to code level, simulation, and implementation). Examples of MoCCs in our case are mainly synchronous reactive formalisms and dataflow process networks. Various extensions or specific restrictions enforce respectively greater expressivity or more focused decidable analysis results.

DataFlow Process Networks and Synchronous Reactive Languages such as ESTEREL/SYNCCHARTS and SIGNAL/POLYCHRONY [53], [54], [48], [15], [4], [13] share one main characteristics: they are specified in a self-timed or loosely timed fashion, in the asynchronous data-flow style. But formal criteria in their semantics ensure that, under good correctness conditions, a sound synchronous interpretation can be provided, in which all treatments (computations, signaling communications) are precisely temporally mapped. This is refered to as clock calculus in synchronous reactive systems, and leads to a large body of theoretical studies and deep results in the case of DataFlow Process Networks [49], [47] (consider SDF balance equations for instance [56]).

As a result, explicit schedules become an important ingredient of design, which ultimately can be considered and handled by the designer him/herself. In practice such schedules are sought to optimize other parts of the design, mainly buffering queues: production and consumption of data can be regulated in their relative speeds. This was specially taken into account in the recent theories of Latency-Insensitive Design [50], or N-synchronous processes [51], with some of our contributions [6].

Explicit schedule patterns should be pictured in the framework of low-power distributed mapping of embedded applications onto manycore architectures, where they could play an important role as theoretical formal models on which to compute and optimize allocations and performances. We describe below two lines of research in this direction. Striking in these techniques is the fact that they include time and timing as integral parts of early functional design. But this original time is logical, multiform, and only partially ordering the various functional computations and communications. This approach was radically generalized in our team to a methodology for logical time based design, described next (see 3.2).

3.1.1. K-periodic static scheduling and routing in Process Networks

In the recent years we focused on the algorithm treatments of ultimately k-periodic schedule regimes, which are the class of schedules obtained by many of the theories described above. An important breakthrough occurred when realizing that the type of ultimatelly periodic binary words that were used for reporting *static scheduling* results could also be employed to record a completely distinct notion of ultimately k-periodic route switching patterns, and furthermore that commonalities of representation could ease combine them together. A new model, by the name of K-periodical Routed marked Graphs (KRG) was introduced, and extensively studied for algebraic and algorithmic properties [5].

The computations of optimized static schedules and other optimal buffering configurations in the context of latency-insensitive design led to the K-Passa software tool development 5.2 .

3.1.2. Endochrony and GALS implementation of conflict-free polychronous programs

The possibility of exploring various schedulings for a given application comes from the fact that some behaviors are truly concurrent, and mutually *conflict-free* (so they can be executed independently, with any choice of ordering). Discovering potential asynchronous inside synchronous reactive specifications then becomes something highly desirable. It can benefit to potential distributed implementation, where signal communications are restricted to a minimum, as they usually incur loss in performance and higher power consumption. This general line of research has come to be known as Endochrony, with some of our contributions [11].

3.2. Logical Time in Model-Driven Embedded System Design

Participants: Julien Deantoni, Frédéric Mallet, Marie Agnès Peraldi Frati, Robert de Simone.

Starting from specific needs and opportunities for formal design of embedded systems as learned from our work on MoCCs (see 3.1), we developed a Logical Time Model as part of the official OMG UML profile MARTE for Modeling and Analysis of Real-Time Embedded systems. With this model is associated a Clock Constraint Specification Language (CCSL), which allows to provide loose or strict logical time constraints between design ingredients, be them computations, communications, or any kind of events whose repetitions can be conceived as generating a logical conceptual clock (or activation condition). The definition of CCSL is provided in [1].

Our vision is that many (if not all) of the timing constraints generally expressed as physical prescriptions in real-time embedded design (such as periodicity, sporadicity) could be expressed in a logical setting, while actually many physical timing values are still unknown or unspecified at this stage. On the other hand, our logical view may express much more, such as loosely stated timing relations based on partial orderings or partial constraints.

So far we have used CCSL to express important phenonema as present in several formalisms: AADL (used in avionics domain), EAST-ADL2 (proposed for the AutoSar automotive electronic design approach), IP-Xact (for System-on-Chip (*SoC*) design). The difference here comes from the fact that these formalisms were formerly describing such issues in informal terms, while CCSL provides a dedicated formal mathematical notation. Close connections with synchronous and polychronous languages, especially Signal, were also established; so was the ability of CCSL to model dataflow process network static scheduling.

In principle the MARTE profile and its Logical Time Model can be used with any UML editor supporting profiles. In practice we focused on the PAPYRUS open-source editor, mainly from CEA LIST. We developed under Eclipse the TIMESQUARE solver and emulator for CCSL constraints (see 5.1), with its own graphical interface, as a stand-alone software module, while strongly coupled with MARTE and Papyrus.

While CCSL constraints may be introduced as part of the intended functionality, some may also be extracted from requirements imposed either from real-time user demands, or from the resource limitations and features from the intended execution platform. Sophisticated detailed descriptions of platform architectures are allowed using MARTE, as well as formal allocations of application operations (computations and communications) onto platform resources (processors and interconnects). This is of course of great value at a time where embedded architectures are becoming more and more heterogeneous and parallel or distributed, so that application mapping in terms of spatial allocation and temporal scheduling becomes harder and harder. This approach is extensively supported by the MARTE profile and its various models. As such it originates from the Application-Architecture-Adequation (AAA) methodology, first proposed by Yves Sorel, member of Aoste. AAA aims at specific distributed real-time algorithmic methods, described next in 3.3 .

Of course, while logical time in design is promoted here, and our works show how many current notions used in real-time and embedded systems synthesis can naturally be phrased in this model, there will be in the end a phase of validation of the logical time assumptions (as is the case in synchronous circuits and SoC design with timing closure issues). This validation is usually conducted from Worst-Case Execution Time (WCET) analysis on individual components, which are then used in further analysis techniques to establish the validity of logical time assumptions (as partial constraints) asserted during the design.

3.3. The AAA (Algorithm-Architecture Adequation) methodology and Real-Time Scheduling

Participants: Laurent George, Dumitru Potop Butucaru, Yves Sorel.

Note: The AAA methodology and the SynDEx environment are fully described at http://www.syndex.org/, together with relevant publications.

3.3.1. Algorithm-Architecture Adequation

The AAA methodology relies on distributed real-time scheduling and relevant optimization to connect an Algorithm/Application model to an Architectural one. We now describe its premises and benefits.

The Algorithm model is an extension of the well known data-flow model from Dennis [52]. It is a directed acyclic hyper-graph (DAG) that we call “conditioned factorized data dependence graph”, whose vertices are “operations” and hyper-edges are directed “data or control dependences” between operations. The data dependences define a partial order on the operations execution. The basic data-flow model was extended in three directions: first infinite (resp. finite) repetition of a sub-graph pattern in order to specify the reactive aspect of real-time systems (resp. in order to specify the finite repetition of a sub-graph consuming different data similar to a loop in imperative languages), second “state” when data dependences are necessary between different infinite repetitions of the sub-graph pattern introducing cycles which must be avoided by introducing specific vertices called “delays” (similar to z^{-n} in automatic control), third “conditioning” of an operation by a control dependence similar to conditional control structure in imperative languages, allowing the execution of alternative subgraphs. Delays combined with conditioning allow the programmer to specify automata necessary for describing “mode changes”.

The Architecture model is a directed graph, whose vertices are of two types: “processor” (one sequencer of operations and possibly several sequencers of communications) and “medium” (support of communications), and whose edges are directed connections.

The resulting implementation model [9] is obtained by an external compositional law, for which the architecture graph operates on the algorithm graph. Thus, the result of such compositional law is an algorithm graph, "architecture-aware", corresponding to refinements of the initial algorithm graph, by computing spatial (distribution) and timing (scheduling) allocations of the operations onto the architecture graph resources. In that context "Adequation" refers to some search amongst the solution space of resulting algorithm graphs, labelled by timing characteristics, for one algorithm graph which verifies timing constraints and optimizes some criteria, usually the total execution time and the number of computing resources (but other criteria may exist). The next section describes distributed real-time schedulability analysis and optimization techniques for that purpose.

3.3.2. Distributed Real-Time Scheduling and Optimization

We address two main issues: uniprocessor and multiprocessor real-time scheduling where constraints must mandatorily be met, otherwise dramatic consequences may occur (hard real-time) and where resources must be minimized because of embedded features.

In the case of uniprocessor real-time scheduling, besides the classical deadline constraint, often equal to a period, we take into consideration dependences beetween tasks and several, latencies. The latter are complex related “end-to-end” constraints. Dealing with multiple real-time constraints raises the complexity of the scheduling problems. Moreover, because the preemption leads, at least, to a waste of resources due to its approximation in the WCET (Worst Execution Time) of every task, as proposed by Liu and Leyland [57], we first studied non-preemtive real-time scheduling with dependences, periodicities, and latencies constraints. Although a bad approximation of the preemption cost, may have dramatic consequences on real-time scheduling, there are only few researches on this topic. We have been investigating preemptive real-time scheduling since few years, and we focus on the exact cost of the preemption. We have integrated this cost in the schedulability conditions that we propose, and in the corresponding scheduling algorithms. More generally, we are interested in integrating in the schedulability analyses the cost of the RTOS (Real-Time Operating

System), for which the cost of preemption is the most difficult part because it varies according to the instance (job) of each task.

In the case of multiprocessor real-time scheduling, we chose at the beginning the partitioned approach, rather than the global approach, since the latter allows task migrations whose cost is prohibitive for current commercial processors. The partitioned approach enables us to reuse the results obtained in the uniprocessor case in order to derive solutions for the multiprocessor case. We consider also the semi-partitioned approach which allows only some migrations in order to minimize the overhead they involve. In addition to satisfy the multiple real-time constraints mentioned in the uniprocessor case, we have to minimize the total execution time (makespan) since we deal with automatic control applications involving feedback loops. Furthermore, the domain of embedded systems leads to solving minimization resources problems. Since these optimization problems are NP-hard we develop exact algorithms (B & B, B & C) which are optimal for simple problems, and heuristics which are sub-optimal for realistic problems corresponding to industrial needs. Long time ago we proposed a very fast "greedy" heuristics [8] whose results were regularly improved, and extended with local neighborhood heuristics, or used as initial solutions for metaheuristics.

In addition to the spatial dimension (distributed) of the real-time scheduling problem, other important dimensions are the type of communication mechanisms (shared memory vs. message passing), or the source of control and synchronization (event-driven vs. time-triggered). We explore real-time scheduling on architectures corresponding to all combinations of the above dimensions. This is of particular impact in application domains such as automotive and avionics (see 4.2).

The arrival of complex hardware responding to the increasing demand for computing power in next generation systems exacerbates the limitations of the current worst-case real-time reasoning. Our solution to overcome these limitations is based on the fact that worst-case situations may have a extremely low probability of appearance within one hour of functioning (10^{-45}), compared to the certification requirements for instance (10^{-9} for the highest level of certification in avionics). Thus we model and analyze the real-time systems using probabilistic models and we propose results that are fundamental for the probabilistic worst-case reasoning over a given time window.

ARIC Project-Team

3. Research Program

3.1. Lattice-based cryptography

Lattice-based cryptography (LBC) is an utterly promising, attractive (and competitive) research ground in cryptography, thanks to a combination of unmatched properties:

- **Improved performance.** LBC primitives have low asymptotic costs, but remain cumbersome in practice (e.g., for parameters achieving security against computations of up to 2100 bit operations). To address this limitation, a whole branch of LBC has evolved where security relies on the restriction of lattice problems to a family of more structured lattices called *ideal lattices*. Primitives based on such lattices can have quasi-optimal costs (i.e., quasi-constant amortized complexities), outperforming all contemporary primitives. This asymptotic performance sometimes translates into practice, as exemplified by NTRUEncrypt.
- **Improved security.** First, lattice problems seem to remain hard even for quantum computers. Moreover, the security of most of LBC holds under the assumption that standard lattice problems are hard in the worst case. Oppositely, contemporary cryptography assumes that specific problems are hard with high probability, for some precise input distributions. Many of these problems were artificially introduced for serving as a security foundation of new primitives.
- **Improved flexibility.** The master primitives (encryption, signature) can all be realized based on worst-case (ideal) lattice assumptions. More evolved primitives such as ID-based encryption (where the public key of a recipient can be publicly derived from its identity) and group signatures, that were the playing-ground of pairing-based cryptography (a subfield of elliptic curve cryptography), can also be realized in the LBC framework, although less efficiently and with restricted security properties. More intriguingly, lattices have enabled long-wished-for primitives. The most notable example is homomorphic encryption, enabling computations on encrypted data. It is the appropriate tool to securely outsource computations, and will help overcome the privacy concerns that are slowing down the rise of the cloud.

We will work on three directions, detailed now.

3.1.1. Lattice algorithms

All known lattice reduction algorithms follow the same design principle: perform a sequence of small elementary steps transforming a current basis of the input lattice, where these steps are driven by the Gram-Schmidt orthogonalisation of the current basis.

In the short term, we will fully exploit this paradigm, and hopefully lower the cost of reduction algorithms with respect to the lattice dimension. We aim at asymptotically fast algorithms with complexity bounds closer to those of basic and normal form problems (matrix multiplication, Hermite normal form). In the same vein, we plan to investigate the parallelism potential of these algorithms.

Our long term goal is to go beyond the current design paradigm, to reach better trade-offs between run-time and shortness of the output bases. To reach this objective, we first plan to strengthen our understanding of the interplay between lattice reduction and numerical linear algebra (how far can we push the idea of working on approximations of a basis?), to assess the necessity of using the Gram-Schmidt orthogonalisation (e.g., to obtain a weakening of LLL-reduction that would work up to some stage, and save computations), and to determine whether working on generating sets can lead to more efficient algorithms than manipulating bases. We will also study algorithms for finding shortest non-zero vectors in lattices, and in particular look for quantum accelerations.

We will implement and distribute all algorithmic improvements, e.g., within the fplll library. We are interested in high performance lattice reduction computations (see application domains below), in particular in connection/continuation with the HPAC ANR project (algebraic computing and high performance consortium).

3.1.2. Lattice-based cryptography

Our long term goal is to demonstrate the superiority of lattice-based cryptography over contemporary public-key cryptographic approaches. For this, we will 1- Strengthen its security foundations, 2- Drastically improve the performance of its primitives, and 3- Show that lattices allow to devise advanced and elaborate primitives.

The practical security foundations will be strengthened by the improved understanding of the limits of lattice reduction algorithms (see last section). On the theoretical side, we plan to attack two major open problems: Are ideal lattices (lattices corresponding to ideals in rings of integers of number fields) computationally as hard to handle as arbitrary lattices? What is the quantum hardness of lattice problems?

Lattice-based primitives involve two types of operations: sampling from discrete Gaussian distributions (with lattice supports), and arithmetic in polynomial rings such as $(\mathbb{Z}/q\mathbb{Z})[x]/(x^n+1)$ with n a power of 2. When such polynomials are used (which is the case in all primitives that have the potential to be practical), then the underlying algorithmic problem that is assumed hard involves ideal lattices. This is why it is crucial to precisely understand the hardness of lattice problems for this family. We will work on improving both types of operations, both in software and in hardware, concentrating on values of q and n providing security. As these problems are very arithmetic in nature, this will naturally be a source of collaboration with the other Themes of the ARIC team.

Our main objective in terms of cryptographic functionality will be to determine the extent to which lattices can help securing cloud services. For example, is there a way for users to delegate computations on their outsourced dataset while minimizing what the server eventually learns about their data? Can servers compute on encrypted data in an efficiently verifiable manner? Can users retrieve their files and query remote databases anonymously provided they hold appropriate credentials? Lattice-based cryptography is the only approach so far that has allowed to make progress into those directions. We will investigate the practicality of the current constructions, the extension of their properties, and the design of more powerful primitives, such as functional encryption (allowing the recipient to learn only a function of the plaintext message). To achieve these goals, we will in particular focus on cryptographic multilinear maps.

This research axis of ARIC is gaining strength thanks to the recruitment of Benoit Libert. We will be particularly interested in the practical and operational impacts, and for this reason we envision a collaboration with an industrial partner.

3.1.3. Application domains

- Diophantine equations. Lattice reduction algorithms can be used to solve diophantine equations, and in particular to find simultaneous rational approximations to real numbers. We plan to investigate the interplay between this algorithmic task, the task of finding integer relations between real numbers, and lattice reduction. A related question is to devise LLL-reduction algorithms that exploit specific shapes of input bases. This will be done within the ANR DynA3S project.
- Communications. We will continue our collaboration with Cong Ling on the use of lattices in communications. We plan to work on the wiretap channel over a fading channel (modeling cell phone communications in a fast moving environment). The current approaches rely on ideal lattices, and we hope to be able to find new approaches thanks to our expertise on them due to their use in lattice-based cryptography. We will also tackle the problem of sampling vectors from Gaussian distributions with lattice support, for a very small standard deviation parameter. This would significantly improve current schemes for communication schemes based on lattices, as well as several cryptographic primitives.
- Cryptanalysis of variants of RSA. Lattices have been used extensively to break variants of the RSA encryption scheme, via Coppersmith's method to find small roots of polynomials. We plan to work with Nadia Heninger (U. of Pennsylvania) on improving these attacks, to make them more practical.

This is an excellent test case for testing the practicality of LLL-type algorithm. Nadia Heninger has a strong experience in large scale cryptanalysis based on Coppersmith's method (http://smartfacts.cr.yp.to/)

3.2. Efficient approximation methods

3.2.1. Computer algebra generation of certified approximations.

We plan to focus on the generation of certified and efficient approximations for solutions of linear differential equations. These functions cover many classical mathematical functions and many more can be built by combining them. One classical target area is the numerical evaluation of elementary or special functions. This is currently performed by code specifically handcrafted for each function. The computation of approximations and the error analysis are major steps of this process that we want to automate, in order to reduce the probability of errors, to allow one to implement "rare functions", to quickly adapt a function library to a new context: new processor, new requirements – either in terms of speed or accuracy.

In order to significantly extend the current range of functions under consideration, several methods originating from approximation theory have to be considered (divergent asymptotic expansions; Chebyshev or generalized Fourier expansions; Padé approximants; fixed point iterations for integral operators). We have done preliminary work on some of them. Our plan is to revisit them all from the points of view of effectivity, computational complexity (exploiting linear differential equations to obtain efficient algorithms), as well as in their ability to produce provable error bounds. This work is to constitute a major progress towards the automatic generation of code for moderate or arbitrary precision evaluation with good efficiency. Other useful, if not critical, applications are certified quadrature, the determination of certified trajectories of spatial objects and many more important questions in optimal control theory.

3.2.2. Digital Signal Processing.

As computer arithmeticians, a wide and important target for us is the design of efficient and certified linear filters in digital signal processing (DSP). Actually, following the advent of Matlab as the major tool for filter design, the DSP experts now systematically delegate to Matlab all the part of the design related to numerical issues. And yet, various key Matlab routines are neither optimized, nor certified. Therefore, there is a lot of room for enhancing numerous DSP numerical implementations and there exist several promising approaches to do so.

The first important challenge that we want to address is the development and the implementation of optimal methods for rounding the coefficients involved in the design of the filter. If done in a naive way, this rounding may lead to a significant loss of performance. We will study in particular FIR and IIR filters.

3.2.3. Table Maker's Dilemma (TMD).

There is a clear demand for hardest-to-round cases, and several computer manufacturers recently contacted us to obtain new cases. These hardest-to-tound cases are a precious help for building libraries of correctly rounded mathematical functions. The current code, based on Lefèvre algorithm, will be rewritten and formal proofs will be done. We plan to use uniform polynomial approximation and diophantine techniques in order to tackle the case of the IEEE quad precision and analytic number theory techniques (exponential sums estimates) for counting the hardest-to-round cases.

3.3. High-performance reliable kernels

The main theme here is the study of fundamental operations ("kernels") on a hierarchy of symbolic or numeric data types spanning integers, floating-point numbers, polynomials, power series, as well as matrices of all these. Fundamental operations include basic arithmetic (e.g., how to multiply or how to invert) common to all such data, as well as more specific ones (change of representation/conversions, GCDs, determinants, etc.). For such operations, which are ubiquitous and at the very core of computing (be it numerical, symbolic, or hybrid numeric-symbolic), our goal is to ensure both high-performance and reliability.

3.3.1. Algorithmic design and analysis of symbolic or numerical algorithms.

On the symbolic side, we have so far obtained fast algorithms for basic operations on both polynomial matrices and structured matrices, but in a rather independent way. Both types turn out to have much in common, but this is sometimes not reflected by the complexities obtained, especially for applications in cryptology and coding theory. Our long term goal in this area is thus to explore these connections further, to provide a more unified treatment and bridge these complexity gaps, and to produce associated efficient implementations. A first step towards this goal will be the design and implementation of enhanced algorithms for various generalizations of Hermite-Padé approximation; in the context of list decoding, this should in particular make it possible to improve over the structured-matrix approach, which is so far the fastest known.

On the numerical side, we will continue to revisit and improve the classical error bounds of numerical analysis in the light of all the subtleties of IEEE floating-point arithmetic. These aspects will be developed jointly with the "symbolic floating-point" approach presented in the next paragraph. A complementary approach will also be studied, based on the estimation (possibly via automatic differentiation) of condition numbers in order to identify inputs leading to large backward errors. Finally, concerning interval arithmetic, a thorough analysis of the accuracy of several representations, such as mid-rad, is also to be done.

3.3.2. Symbolic floating-point arithmetic.

Our work on the analysis of algorithms in floating-point arithmetic leads us to manipulate floating-point data in their greatest generality, that is, as symbolic expressions in the base and the precision. A long-term goal here is to develop theorems as well as efficient data structures and algorithms for handling such quantities by computer rather than by hand as we do now. This is a completely new direction, whose main outcome will be a "symbolic floating-point toolbox" distributed in computer algebra systems like Sage and or Maple. In particular, such a toolbox will provide a way to check automatically the certificates of optimality we have obtained on the error bounds of various numerical algorithms. A PhD student has started on this subject in September 2014.

3.3.3. High-performance multiple precision arithmetic libraries.

Many numerical problems require higher precision than the conventional floating-point (single, double) formats. One solution is to use multiple precision libraries such as GNU MPFR, which allow the manipulation of very high precision numbers, but their generality (they are able to handle numbers with millions of digits), is a quite heavy alternative when high performance is needed. Our objective is to design a multiple precision arithmetic library that would allow to tackle problems where a precision of a few hundred bits is sufficient, but which have strong performance requirements. Applications include the process of long-term iteration of chaotic dynamical systems ranging from the classical Henon map to calculations of planetary orbits. The designed algorithms will be formally proved. We are in close contact with Warwick Tucker (Uppsala University, Sweden) and Mioara Joldes (LAAS, Toulouse) on this topic. A PhD student funded by a Région Rhône-Alpes grant has started on this topic in September 2014.

3.3.4. Interactions between arithmetics.

We will work on the interplay between floating-point and integer arithmetics, and especially on how to make the best use of both integer and floating-point basic operations when designing floating-point numerical kernels for embedded devices. This will be done in the context of the Metalibm ANR project and of our collaboration with STMicroelectronics. In addition, our work on the IEEE 1788 standard leads naturally to the development of associated reference libraries for interval arithmetic. A first direction will be to implement IEEE 1788 interval arithmetic using the fixed-precision hardware available for IEEE 754-2008 floating-point arithmetic. Another one will be to provide efficient support for multiple-precision intervals, in mid-rad representation and by developing MPFR-based code-generation tools aimed at handling families of functions.

3.3.5. Adequation algorithms/architectures.

So far, we have investigated how specific instructions like the fused multiply-add (FMA) impact the accuracy of computations, and have proposed several highly accurate FMA-based algorithms. The FMA being available

on several recent architectures, we now want to understand its impact on such algorithms in terms of practical performances. This should be a medium term project, leading to FMA-based algorithms with best speed/accuracy/robustness tradeoff. On the other hand (and on the long term), a major issue is how to exploit the various levels of parallelism of recent and upcoming architectures to ensure simultaneously high performance and reliability. A first direction will be to focus on SIMD parallelism, offered by instruction sets via vector instructions. This kind of parallelism should be key for small numerical kernels like elementary functions, complex arithmetic, or low-dimensional matrix computations. A second direction will be at the multi-core processor level, especially for larger numerical or algebraic problems (and in conjunction with SIMD parallelism when handling sub-problems of small enough dimension). Finally, we will work on aspects of automatic adaptation (auto-tuning) to such architectural features, not only for speed, but also for accuracy. This could be done via the design and implementation of heuristics capable of inserting more accurate codes, based for example on error-free transforms, whenever needed.

ATEAMS Project-Team

3. Research Program

3.1. Research method

We are inspired by formal methods and logic to construct new tools for software analysis, transformation and generation. We try and proof the correctness of new algorithms using any means necessary.

Nevertheless we mainly focus on the study of existing (large) software artifacts to validate the effectiveness of new tools. We apply the scientific method. To (in)validate our hypothesis we often use detailed manual source code analysis, or we use software metrics, and we have started to use more human subjects (programmers).

Note that we maintain ties with the CWI spinoff "Software Improvement Group" which services most of the Dutch software industry and government and many European companies as well. This provides access to software systems and information about software systems that is valuable in our research.

3.2. Software analysis

This research focuses on source code; to analyze it, transform it and generate it. Each analysis or transformation begins with fact extraction. After that we may analyze specific software systems or large bodies of software systems. Our goal is to improve software systems by understanding and resolving the causes of software complexity. The approach is captured in the EASY acronym: Extract Analyze SYnthesize. The first step is to extract facts from source code. These facts are then enriched and refined in an analysis phase. Finally the result is synthesized in the form of transformed or generated source code, a metrics report, a visualization or some other output artifact.

The mother and father of fact extraction techniques are probably Lex, a scanner generator, and AWK, a language intended for fact extraction from textual records and report generation. Lex is intended to read a file character-by-character and produce output when certain regular expressions (for identifiers, floating point constants, keywords) are recognized. AWK reads its input line-by-line and regular expression matches are applied to each line to extract facts. User-defined actions (in particular print statements) can be associated with each successful match. This approach based on regular expressions is in wide use for solving many problems such as data collection, data mining, fact extraction, consistency checking, and system administration. This same approach is used in languages like Perl, Python, and Ruby. Murphy and Notkin have specialized the AWK-approach for the domain of fact extraction from source code. The key idea is to extend the expressivity of regular expressions by adding context information, in such a way that, for instance, the begin and end of a procedure declaration can be recognized. This approach has, for instance, been used for call graph extraction but becomes cumbersome when more complex context information has to be taken into account such as scope information, variable qualification, or nested language constructs. This suggests using grammar-based approaches as will be pursued in the proposed project. Another line of research is the explicit instrumentation of existing compilers with fact extraction capabilities. Examples are: the GNU C compiler GCC, the CPPX C++ compiler, and the Columbus C/C++ analysis framework. The Rigi system provides several fixed fact extractors for a number of languages. The extracted facts are represented as tuples (see below). The CodeSurfer source code analysis tool extracts a standard collection of facts that can be further analyzed with built-in tools or user-defined programs written in Scheme. In all these cases the programming language as well as the set of extracted facts are fixed thus limiting the range of problems that can be solved.

The approach we are exploring is the use of syntax-related program patterns for fact extraction. An early proposal for such a pattern-based approach consisted of extending a fixed base language (either C or PL/1 variant) with pattern matching primitives. In our own previous work on RScript we have already proposed a query algebra to express direct queries on the syntax tree. It also allows the querying of information that is attached to the syntax tree via annotations. A unifying view is to consider the syntax tree itself as "facts" and to represent it as a relation. This idea is already quite old. For instance, Linton proposes to represent all syntactic as well as semantic aspects of a program as relations and to use SQL to query them. Due to the lack of expressiveness of SQL (notably the lack of transitive closure) and the performance problems encountered, this approach has not seen wider use.

Parsing is a fundamental tool for fact extraction for source code. Our group has longstanding contributions in the field of Generalized LR parsing and Scannerless parsing. Such generalized parsing techniques enable generation of parsers for a wide range of existing (legacy) programming languages, which is highly relevant for experimental research and validation.

Extracted facts are often refined, enriched and queried in the analysis phase. We propose to use a relational formalization of the facts. That is, facts are represented as sets of tuples, which can then be queried using relational algebra operators (e.g., domain, transitive closure, projection, composition etc.). This relational representation facilitates dealing with graphs, which are commonly needed during program analysis, for instance when processing control-flow or data-flow graphs. The Rascal language integrates a relational sub-language by providing comprehensions over different kinds of data types, in combination with powerful pattern matching and built-in primitives for computing (transitive/reflexive) closures and fixpoint computations (equation solving).

3.2.1. Goals

The main goal is to replace labour-intensive manual programming of fact extractors by automatic generation based on concise and formal specification. There is a wide open scientific challenge here: to create a uniform and generic framework for fact extraction that is superior to current more ad-hoc approaches, yet flexible enough to be customized to the analysis case at hand. We expect to develop new ideas and techniques for generic (language-parametric) fact extraction from source code and other software artifacts.

Given the advances made in fact extraction we are starting to apply our techniques to observe source code and analyze it in detail.

3.3. Refactoring and Transformation

The second goal, to be able to safely refactor or transform source code can be realized in strong collaboration with extraction and analysis.

Software refactoring is usually understood as changing software with the purpose of increasing its readability and maintainability rather than changing its external behavior. Refactoring is an essential tool in all agile software engineering methodologies. Refactoring is usually supported by an interactive refactoring tool and consists of the following steps:

- Select a code fragment to refactor.
- Select a refactoring to apply to it.
- Optionally, provide extra parameter needed by the refactoring (e.g., a new name in a renaming).

The refactoring tool will now test whether the preconditions for the refactoring are satisfied. Note that this requires fact extraction from the source code. If this fails the user is informed. The refactoring tool shows the effects of the refactoring before effectuating them. This gives the user the opportunity to disable the refactoring in specific cases.The refactoring tool applies the refactoring for all enabled cases. Note that this implies a transformation of the source code. Some refactorings can be applied to any programming language (e.g., rename) and others are language specific (e.g., Pull Up Method). At http://www.refactoring.com an extensive list of refactorings can be found.

There is hardly any general and pragmatic theory for refactoring, since each refactoring requires different static analysis techniques to be able to check the preconditions. Full blown semantic specification of programming languages have turned out to be infeasible, let alone easily adaptable to small changes in language semantics. On the other hand, each refactoring is an instance of the extract, analyze and transform paradigm. Software transformation regards more general changes such as adding functionality and improving non-functional properties like performance and reliability. It also includes transformation from/to the same language (source-to-source translation) and transformation between different languages (conversion, translation). The underlying techniques for refactoring and transformation are mostly the same. We base our source code transformation techniques on the classical concept of term rewriting, or aspects thereof. It offers simple but powerful pattern matching and pattern construction features (list matching, AC Matching), and type-safe heterogenous data-structure traversal methods that are certainly applicable for source code transformation.

3.3.1. Goals

Our goal is to integrate the techniques from program transformation completely with relational queries. Refactoring and transformation form the Achilles Heel of any effort to change and improve software. Our innovation is in the strict language-parametric approach that may yield a library of generic analyses and transformations that can be reused across a wide range of programming and application languages. The challenge is to make this approach scale to large bodies of source code and rapid response times for precondition checking.

3.4. The Rascal Meta-programming language

The Rascal Domain-Specific Language for Source code analysis and Transformation is developed by ATeams. It is a language specifically designed for any kind of meta programming.

Meta programming is a large and diverse area both conceptually and technologically. There are plentiful libraries, tools and languages available but integrated facilities that combine both source code analysis and source code transformation are scarce. Both domains depend on a wide range of concepts such as grammars and parsing, abstract syntax trees, pattern matching, generalized tree traversal, constraint solving, type inference, high fidelity transformations, slicing, abstract interpretation, model checking, and abstract state machines. Examples of tools that implement some of these concepts are ANTLR, ASF+SDF, CodeSurfer, Crocopat, DMS, Grok, Stratego, TOM and TXL. These tools either specialize in analysis or in transformation, but not in both. As a result, combinations of analysis and transformation tools are used to get the job done. For instance, ASF+SDF relies on RScript for querying and TXL interfaces with databases or query tools. In other approaches, analysis and transformation are implemented from scratch, as done in the Eclipse JDT. The TOM tool adds transformation primitives to Java, such that libraries for analysis can be used directly. In either approach, the job of integrating analysis with transformation has to be done over and over again for each application and this requires a significant investment.

We propose a more radical solution by completely merging the set of concepts for analysis and transformation of source code into a single language called Rascal. This language covers the range of applications from pure analyses to pure transformations and everything in between. Our contribution does not consist of new concepts or language features *per se*, but rather the careful collaboration, integration and cross-fertilization of existing concepts and language features.

3.4.1. Goals

The goals of Rascal are: (a) to remove the cognitive and computational overhead of integrating analysis and transformation tools, (b) to provide a safe and interactive environment for constructing and experimenting with large and complicated source code analyses and transformations such as, for instance, needed for refactorings, and (c) to be easily understandable by a large group of computer programming experts. Rascal is not limited to one particular object programming language, but is generically applicable. Reusable, language specific, functionality is realized as libraries. As an end-result we envision Rascal to be a one-stop shop for source code analysis, transformation, generation and visualization.

3.5. Domain-specific Languages

Our final goal is centered around Domain-specific languages (DSLs), which are software languages tailored to a specific problem domain. DSLs can provide orders of magnitude improvement in terms of software quality and productivity. However, the implementation of DSLs is challenging and requires not only thorough knowledge of the problem domain (e.g., finance, digital forensics, insurance, auditing etc.), but also knowledge of language implementation (e.g., parsing, compilation, type checking etc.). Tools for language implementation have been around since the archetypical parser generator YACC. However, many of such tools are characterized by high learning curves, lack of integration of language implementation facets, and lead to implementations that are hard to maintain. This line of research focuses on two topics: improve the practice and experience of DSL implementation, and evaluate the success of DSLs in industrial practice.

Language workbenches [4] are integrated environments to facilitate the development of all aspects of DSLs. This includes IDE support (e.g., syntax coloring, outlining, reference resolving etc.) for the defined languages. Rascal can be seen as a language workbench that focuses on flexibility, programmability and modularity. DSL implementation is, in essence, an instance of source code analysis and transformation. As a result, Rascal's features for fact extraction, analysis, tree traversal and synthesis are an excellent fit for this area. An important aspect in this line of research is bringing the IDE closer to the source code. This will involve investigation of heterogeneous representations of source code, by integrating graphical, tabular or forms-based user interface elements. As a result, we propose Rascal as a feature-rich workbench for model-driven software development.

The second component of this research is concerned with evaluating DSLs in industrial contexts. This means that DSLs constructed using Rascal will be applied in real-life environments so that expected improvements in quality, performance, or productivity can be observed. We already have experience with this in the domain of digital forensics, computational auditing and games.

3.5.1. Goals

The goal of this research topic is to improve the practice of DSL-based software development through language design and tool support. A primary focus is to extend the IDE support provided by Rascal, and to facilitate incremental, and iterative design of DSLs. The latter is supported by new (meta-)language constructs for extending existing language implementations. This will require research into extensible programming and composition of compilers, interpreters and type checkers. Finally, a DSL is never an island: it will have to integrate with (third-party) source code, such as host language, libraries, runtime systems etc. This leads to the vision of multi-lingual programming environments [15].

CAIRN Project-Team

3. Research Program

3.1. Panorama

The development of complex applications is traditionally split in three stages: a theoretical study of the algorithms, an analysis of the target architecture and the implementation. When facing new emerging applications such as high-performance, low-power and low-cost mobile communication systems or smart sensor-based systems, it is mandatory to strengthen the design flow by a joint study of both algorithmic and architectural issues [0].

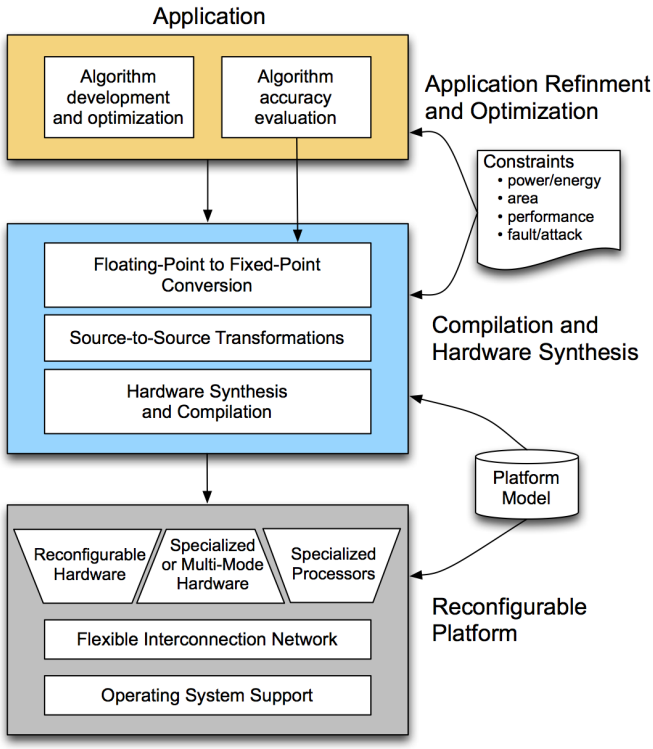


Figure 1. CAIRN*'s general design flow and related research themes*

Figure 1 shows the global design flow we propose to develop. This flow is organized in levels which refer to our three research themes: application optimization (new algorithms, fixed-point arithmetic and advanced representations of numbers), architecture optimization (reconfigurable and specialized hardware, application-specific processors), and stepwise refinement and code generation (code transformations, hardware synthesis, compilation).

[0] Often referenced as algorithm-architecture mapping or interaction.

In the rest of this part, we briefly describe the challenges concerning **new reconfigurable platforms** in Section 3.2 , the issues on **compiler and synthesis tools** related to these platforms in Section 3.3 , and the remaining challenges in **algorithm architecture interaction** in Section 3.4 .

3.2. Reconfigurable Architecture Design

Over the last two decades, there has been a strong push of the research community to evolve static programmable processors into run-time dynamic and partial reconfigurable (DPR) architectures. Several research groups around the world have hence proposed reconfigurable hardware systems operating at various levels of granularity. For example, functional-level reconfiguration has been proposed to increase the efficiency of programmable processors without having to pay for the FPGAs penalties. These coarse-grained reconfigurable architectures (CGRAs) provide operator-level configurable functional blocks and word-level datapaths. The main goal of this class of architectures is to provide flexibility while minimizing reconfiguration overhead (there exists several recent surveys on this topic [113], [97], [78], [114]). Compared to fine-grained architectures, CGRAs benefit from a massive reduction in configuration memory and configuration delay, as well as a considerable reduction in routing and placement complexity. This, in turns, results in an improvement in the computation volume over energy cost ratio, even if it comes at the price of a loss of flexibility compared to bit-level operations. Such constraints have been taken into account in the design of DART [93][12], CRIP [81], Adres [105] or others [116]. These works have led to commercial products such as the Extreme Processor Platform (XPP) [82] from PACT or Montium [0] from Recore systems.

Another strong trend is the design of hybrid architectures which combine standard GPP or DSP cores with arrays of *configurable elements* such as the Lx [96], or of *field-configurable elements* such as the Xirisc processor [103] and more recently by commercial platforms such as the Xilinx Zynq-7000. Some of their benefits are the following: functionality on demand (set-top boxes for digital TV equipped with decoding hardware on demand), acceleration on demand (coprocessors that accelerate computationally demanding applications in multimedia or communications applications), and shorter time-to-market (products that target ASIC platforms can be released earlier using reconfigurable hardware).
Dynamic reconfiguration enables an architecture to adapt itself to various incoming tasks. This requires complex resource management and control which can be provided as services by a real-time operating system (RTOS) [104]: communication, memory management, task scheduling [92], [85][1] and task placement. Such an Operating System (OS) based approach has many advantages: it provides a complete design framework, that is independent of the technology and of the underlying hardware architecture, helping to drastically reduce the full platform design time. Due to the unpredictable execution of tasks, the OS must be able to allocate resource to tasks at run-time along with mechanisms to support inter-task communication. An efficient way to support such communications is to resort to a network-on-chip [111]. The role of the communication infrastructure is then to support transactions between different components of the platform, either between macro-components – main processor, dedicated modules, dynamically reconfigurable component – or within the elements of the reconfigurable components themselves.

In CAIRNwe mainly target reconfigurable system-on-chip (RSoC) defined as a set of computing and storing resources organized around a flexible interconnection network and integrated within a single silicon chip (or programmable chip such as FPGAs). The architecture is customized for an application domain, and the flexibility is provided by both hardware reconfiguration and software programmability. Computing resources are therefore highly heterogeneous and raise many issues that we discuss in the following:

- **Reconfigurable hardware blocks with a dynamic behavior** where reconfigurability can be achieved at the bit- or operator-level. Our research aims at defining new reconfigurable architectures including computing and memory resources. Since reconfiguration must happen as fast as possible (typically within a few cycles), reducing the configuration time overhead is also a key issue.

[0] http://www.recoresystems.com/

- When performance and power consumption are major constraints, it is acknowledged that optimized specialized hardware blocks (often called IPs for Intellectual Properties) are the best (and often the only) solution. Therefore, we also study architecture and tools for **specialized hardware accelerators** and for **multi-mode components**.
- Customized **processors with a specialized instruction-set** also offer a viable solution to trade between energy efficiency and flexibility. They are particularly relevant for modern FPGA platforms where many processor cores can be embedded. For this topic, we focus on the automatic generation of heterogeneous (sequential or parallel) reconfigurable processor extensions that are tightly coupled to processor cores.

3.3. Compilation and Synthesis for Reconfigurable Platforms

In spite of their advantages, reconfigurable architectures lack efficient and standardized compilation and design tools. As of today, this still makes the technology impractical for large scale industrial use. Generating and optimizing the mapping from high-level specifications to reconfigurable hardware platforms is therefore a key research issue, and the problem has received considerable interest over the last years [108], [84], [115], [118]. In the meantime, the complexity (and heterogeneity) of these platforms has also been increasing quite significantly, with complex heterogeneous multi-cores architectures becoming a *de facto* standard. As a consequence, the focus of designers is now geared toward optimizing overall system-level performance and efficiency [99], [108], [107]. Here again, existing tools are not well suited, as they fail at providing a unified programming view of the programmable and/or reconfigurable components implemented on the platform.
In this context we have been pursuing our efforts to propose tools whose design principles are based on a tight coupling between the compiler and the target hardware architectures. We build on the expertise of the team members in High Level Synthesis (HLS) [8], ASIP optimizing compilers [15] and automatic parallelization for massively parallel specialized circuits [6]. We first study how to increase the efficiency of standard programmable processors by extending their instruction set to speed-up computationally-intensive kernels. Our focus is on efficient and exact algorithms for the identification, selection and scheduling of such instructions [9]. We also propose techniques to synthesize reconfigurable (or multi-mode) architectures. We address these challenges by borrowing techniques from high-level synthesis, optimizing compilers and automatic parallelization, especially when dealing with nested loop kernels. The goal is then either to derive a custom fine-grain parallel architecture and/or to derive the configuration of a Coarse Grain Reconfigurable Architecture (CGRA). In addition, and independently of the scientific challenges mentioned above, proposing such flows also poses significant software engineering issues. As a consequence, we also study how leading edge Object Oriented software engineering techniques (Model Driven Engineering) can help the Computer Aided Design (CAD) and optimizing compiler communities prototyping new research ideas.
Efficient implementation of multimedia and signal processing applications (in software for DSP cores or as special-purpose hardware) often requires, for reasons related to cost, power consumption or silicon area constraints, the use of fixed-point arithmetic, whereas the algorithms are usually specified in floating-point arithmetic. Unfortunately, fixed-point conversion is very challenging and time-consuming, typically demanding up to 50% of the total design or implementation time [86]. Thus, tools are required to automate this conversion. For hardware or software implementation, the aim is to optimize the fixed-point specification. The implementation cost is minimized under a numerical accuracy or an application performance constraint. For DSP-software implementation, methodologies have been proposed [101], [106] to achieve a conversion leading to an ANSI-C code with integer data types. For hardware implementation, the best results are obtained when the word-length optimization process is coupled with the high-level synthesis [100], [89]. Evaluating the effects of finite precision is one of the major and often the most time consuming step while performing fixed-point refinement. Indeed, in the word-length optimization process, the numerical accuracy is evaluated as soon as a new word-length is tested, thus, several times per iteration of the optimization process. Classical approaches are based on fixed-point simulations [90], [112]. They lead to long evaluation times and cannot be used to explore the entire design space. Therefore, our aim is to propose closed-form expressions of errors due to fixed-point approximations that are used by a fast analytical framework for accuracy evaluation.

3.4. Interaction between Algorithms and Architectures

As CAIRN mainly targets domain-specific system-on-chip including reconfigurable capabilities, algorithmic-level optimizations have a great potential on the efficiency of the overall system. Based on the skills and experiences in "signal processing and communications" of some CAIRN's members, we conduct research on algorithmic optimization techniques under two main constraints: energy consumption and computation accuracy; and for two main application domains: fourth-generation (4G) mobile communications and wireless sensor networks (WSN). These application domains are very conducive to our research activities. The high complexity of the first one and the stringent power constraint of the second one, require the design of specific high-performance and energy-efficient SoCs. We also consider other applications such as video or bioinformatics, but this short state-of-the-art will be limited to wireless applications.

The radio in both transmit and receive modes consumes the bulk of the total power consumption of the system. Therefore, protocol optimization is one of the main sources of significant energy reduction to be able to achieve self-powered autonomous systems. Reducing power due to radio communications can be achieved by two complementary main objectives: (i) minimizing the output transmit power while maintaining sufficient wireless link quality and (ii) minimizing useless wake-up and channel hearing while still being reactive.

As the physical layer affects all higher layers in the protocol stack, it plays an important role in the energy-constrained design of WSNs. The question to answer can be summarized as: *how much signal processing can be added to decrease the transmission energy (i.e. the output power level at the antenna) such that the global energy consumption be decreased?* The temporal and spatial diversity of relay and multiple antenna techniques are very attractive due to their simplicity and their performance for wireless transmission over fading channels. Cooperative MIMO (multiple-input and multiple-output) techniques have been first studied in [94], [102] and have shown their efficiency in terms of energy consumption [91]. Our research aims at finding new energy-efficient cooperative protocols associating distributed MIMO with opportunistic and/or multiple relays and considering wireless channel impairments such as transmitters desynchronisation.

Another way to reduce the energy consumption consists in decreasing the radio activity, controlled by the medium access (MAC) layer protocols. In this regard, low duty-cycle protocols, such as preamble-sampling MAC protocols, are very efficient because they improve the lifetime of the network by reducing the unnecessary energy waste [80]. As the network parameters (data rate, topology, etc.) can vary, we propose new adaptive MAC protocols to avoid overhearing and idle listening.

Finally, MIMO precoding is now recognized as a very interesting technique to enhance the data rate in wireless systems, and is already used in Wi-Max standard (802.16e). This technique can also be used to reduce transmission energy for the same transmission reliability and the same throughput requirement. One of the most efficient precoders is based on the maximization of the minimum Euclidean distance ($\max$-d_{min}) between two received data vectors [87], but it is difficult to define the closed-form of the optimized precoding matrix for large MIMO system with high-order modulations. Our goal is to derive new generic precoders with simple expressions depending only on the channel angle and the modulation order.

CAMUS Team

3. Research Program

3.1. Research directions

The various objectives we are expecting to reach are directly related to the search of adequacy between the sofware and the new multicore processors evolution. They also correspond to the main research directions suggested by Hall, Padua and Pingali in [28]. Performance, correction and productivity must be the users' perceived effects. They will be the consequences of research works dealing with the following issues:

- Issue 1: Static parallelization and optimization
- Issue 2: Profiling and execution behavior modeling
- Issue 3: Dynamic program parallelization and optimization, virtual machine
- Issue 4: Object-oriented programming and compiling for multicores
- Issue 5: Proof of program transformations for multicores

Efficient and correct applications development for multicore processors needs stepping in every application development phase, from the initial conception to the final run.

Upstream, all potential parallelism of the application has to be exhibited. Here static analysis and transformation approaches (issue 1) must be processed, resulting in a *multi-parallel* intermediate code advising the running virtual machine about all the parallelism that can be taken advantage of. However the compiler does not have much knowledge about the execution environment. It obviously knows the instruction set, it can be aware of the number of available cores, but it does not know the effective available resources at any time during the execution (memory, number of free cores, etc.).

That is the reason why a "virtual machine" mechanism will have to adapt the application to the resources (issue 3). Moreover the compiler will be able to take advantage only of a part of the parallelism induced by the application. Indeed some program information (variables values, accessed memory adresses, etc.) being available only at runtime, another part of the available parallelism will have to be generated on-the-fly during the execution, here also, thanks to a dynamic mechanism.

This on-the-fly parallelism extraction will be performed using speculative behavior models (issue 2), such models allowing to generate speculative parallel code (issue 3). Between our behavior modeling objectives, we can add the behavior monitoring, or profiling, of a program version. Indeed current and future architectures complexity avoids assuming an optimal behavior regarding a given program version. A monitoring process will allow to select on-the-fly the best parallelization.

These different parallelizing steps are schematized on figure 1 .

The more and more widespread usage of object-oriented approaches and languages emphasizes the need for specific multicore programming tools. The object and method formalism implies specific execution schemes that translate in the final binary by quite distant elementary schemes. Hence the execution behavior control is far more difficult. Analysis and optimization, either static or dynamic, must take into account from the outset this distortion between object-oriented specification and final binary code: how can object or method parallelization be translated (issue 4).

Our project lies on the conception of a production chain for efficient execution of an application on a multicore architecture. Each link of this chain has to be formally verified in order to ensure correction as well as efficiency. More precisely, it has to be ensured that the compiler produces a correct intermediate code, and that the virtual machine actually performs the parallel execution semantically equivalent to the source code: every transformation applied to the application, either statically by the compiler or dynamically by the virtual machine, must preserve the initial semantics. They must be proved formally (issue 5).

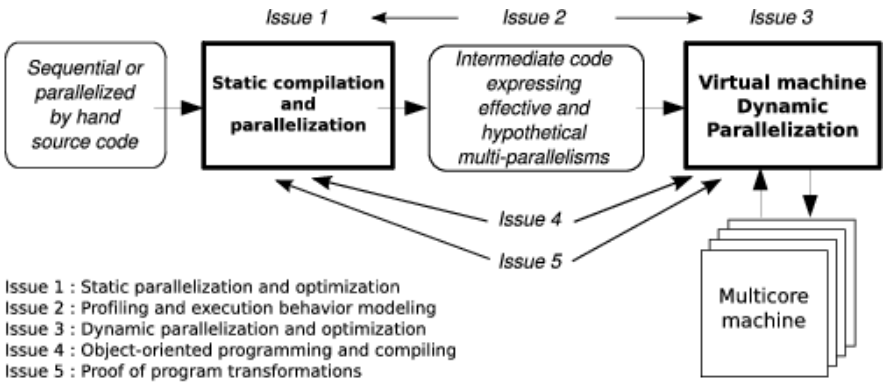


Figure 1. Automatic parallelizing steps for multicore architectures

In the following, those different issues are detailed while forming our global and long term vision of what has to be done.

3.2. Static parallelization and optimization

Participants: Vincent Loechner, Philippe Clauss, Éric Violard, Jean-François Dollinger, Aravind Sukumaran-Rajam, Juan Manuel Martinez Caamaño.

Static optimizations, from source code at compile time, benefit from two decades of research in automatic parallelization: many works address the parallelization of loop nests accessing multi-dimensional arrays, and these works are now mature enough to generate efficient parallel code [27]. Low-level optimizations, in the assembly code generated by the compiler, have also been extensively dealt for single-core and require few adaptations to support multicore architectures. Concerning multicore specific parallelization, we propose to explore two research directions to take full advantage of these architectures: adapting parallelization to multicore architecture and expressing many potential parallelisms.

3.3. Profiling and execution behavior modeling

Participants: Alain Ketterlin, Philippe Clauss, Aravind Sukumaran-Rajam.

The increasing complexity of programs and hardware architectures makes it ever harder to characterize beforehand a given program's run time behavior. The sophistication of current compilers and the variety of transformations they are able to apply cannot hide their intrinsic limitations. As new abstractions like transactional memories appear, the dynamic behavior of a program strongly conditions its observed performance. All these reasons explain why empirical studies of sequential and parallel program executions have been considered increasingly relevant. Such studies aim at characterizing various facets of one or several program runs, *e.g.*, memory behavior, execution phases, etc. In some cases, such studies characterize more the compiler than the program itself. These works are of tremendous importance to highlight all aspects that escape static analysis, even though their results may have a narrow scope, due to the possible incompleteness of their input data sets.

3.4. Dynamic parallelization and optimization, virtual machine

Participants: Aravind Sukumaran-Rajam, Juan Manuel Martinez Caamaño, Jean-François Dollinger, Alexandra Jimborean, Philippe Clauss, Vincent Loechner, Alain Ketterlin.

This link in the programming chain has become essential with the advent of the new multicore architectures. Still being considered as secondary with mono-core architectures, dynamic analysis and optimization are now one of the keys for controling those new mechanisms complexity. From now on, performed instructions are not only dedicated to the application functionalities, but also to its control and its transformation, and so in its own interest. Behaving like a computer virus, such a process should rather be qualified as a “vitamin”. It perfectly knows the current characteristics of the execution environment and owns some qualitative information thanks to a behavior modeling process (issue 2). It appends a significant part of optimizing ability compared to a static compiler, while observing live resources availability evolution.

3.5. Proof of program transformations for multicores

Participants: Éric Violard, Julien Narboux, Nicolas Magaud.

Our main objective consists in certifying the critical modules of our optimization tools (the compiler and the virtual machine). First we will prove the main loop transformation algorithms which constitute the core of our system.

The optimization process can be separated into two stages: the transformations consisting in optimizing the sequential code and in exhibiting parallelism, and those consisting in optimizing the parallel code itself. The first category of optimizations can be proved within a sequential semantics. For the other optimizations, we need to work within a concurrent semantics. We expect the first stage of optimizations to produce data-race free code. For the second stage of optimizations, we will first assume that the input code is data-race free. We will prove those transformations using Appel’s concurrent separation logic [29]. Proving transformations involving program which are not data-race free will constitute a longer term research goal.

CARAMEL Project-Team

3. Research Program

3.1. Cryptography, Arithmetic: Hardware and Software

One of the main topics for our project is public-key cryptography. After 20 years of hegemony, the classical public-key algorithms (whose security is based on integer factorization or discrete logarithm in finite fields) are currently being overtaken by elliptic curves. The fundamental reason for this is that the best algorithms known for factoring integers or for computing discrete logarithms in finite fields have — at best — a subexponential complexity, whereas the best attack known for elliptic-curve discrete logarithms has exponential complexity. As a consequence, for a given security level 2^n, the key sizes must grow linearly with n for elliptic curves, whereas they grow like n^3 for RSA-like systems. As a consequence, several governmental agencies, like the NSA (National Security Agency, USA) or the BSI (Bundesamt für Sicherheit in der Informationstechnik, Germany), now recommend to use elliptic-curve cryptosystems for new products that are not bound to RSA for backward compatibility.

Besides RSA and elliptic curves, there are several alternatives currently under study. There is a recent trend to promote alternate solutions that do not rely on number theory, with the objective of building systems that would resist a quantum computer (in contrast, integer factorization and discrete logarithms in finite fields and elliptic curves have a polynomial-time quantum solution). Among them, we find systems based on hard problems in lattices (NTRU is the most famous), those based on coding theory (McEliece system and improved versions), and those based on the difficulty to solve multivariate polynomial equations (UOV, for instance). None of them has yet reached the same level of popularity as RSA or elliptic curves for various reasons, including the presence of unsatisfactory features (like a huge public key), or the non-maturity (system still alternating between being fixed one day and broken the next day).

Returning to number theory, an alternative to RSA and elliptic curves is to use other curves and in particular genus-2 curves. These so-called hyperelliptic cryptosystems have been proposed in 1989 [32], soon after the elliptic ones, but their deployment is by far more difficult. The first problem was the group law. For elliptic curves, the elements of the group are just the points of the curve. In a hyperelliptic cryptosystem, the elements of the group are points on a 2-dimensional variety associated to the genus-2 curve, called the Jacobian variety. Although there exist polynomial-time methods to represent and compute with them, it took some time before getting a group law that could compete with the elliptic one in terms of speed. Another question that is still not yet fully answered is the computation of the group order, which is important for assessing the security of the associated cryptosystem. This amounts to counting the points of the curve that are defined over the base field or over an extension, and therefore this general question is called point-counting. In the past ten years there have been major improvements on the topic, but there are still cases for which no practical solution is known.

Another recent discovery in public-key cryptography is the fact that having an efficient bilinear map that is hard to invert (in a sense that can be made precise) can lead to powerful cryptographic primitives. The only examples we know of such bilinear maps are associated with algebraic curves, and in particular elliptic curves: this is the so-called Weil pairing (or its variant, the Tate pairing). Initially considered as a threat for elliptic-curve cryptography, they have proven to be quite useful from a constructive point of view, and since the beginning of the decade, hundreds of articles have been published, proposing efficient protocols based on pairings. A long-lasting open question, namely the construction of a practical identity-based encryption scheme, has been solved this way. The first standardization of pairing-based cryptography has recently occurred (see ISO/IEC 14888-3 or IEEE P1363.3), but the recent progress in discrete logarithms in finite fields will probably slow down its large deployment.

Despite the rise of elliptic curve cryptography and the variety of more or less mature alternatives, classical systems (based on factoring or discrete logarithm in finite fields) are still going to be widely used in the next decade, at least, due to resilience: it takes a long time to adopt new standards, and then an even longer time to renew all the software and hardware that is widely deployed.

This context of public-key cryptography motivates us to work on integer factorization, for which we have acquired expertise, both in factoring moderate-sized numbers, using the ECM (Elliptic Curve Method) algorithm, and in factoring large RSA-like numbers, using the number field sieve algorithm. The goal is to follow the transition from RSA to other systems and continuously assess its security to adjust key sizes. We also work on the discrete-logarithm problem in finite fields. This second task is not only necessary for assessing the security of classical public-key algorithms, but is also crucial for the security of pairing-based cryptography.

Another general application for the project is computer algebra systems (CAS), that rely in many places on efficient arithmetic. Nowadays, the objective of a CAS is not only to support an increasing number of features that the user might wish, but also to compute the results fast enough, since in many cases, the CAS are used interactively, and a human is waiting for the computation to complete. To tackle this question, more and more CAS use external libraries, that have been written with speed and reliability as first concern. For instance, most of today's CAS use the GMP library for their computations with big integers. Many of them will also use some external Basic Linear Algebra Subprograms (BLAS) implementation for their needs in numerical linear algebra.

During a typical CAS session, the libraries are called with objects whose sizes vary a lot; therefore being fast on all sizes is important. This encompasses small-sized data, like elements of the finite fields used in cryptographic applications, and larger structures, for which asymptotically fast algorithms are to be used. For instance, the user might want to study an elliptic curve over the rationals, and as a consequence, check its behaviour when reduced modulo many small primes; and then [s]he can search for large torsion points over an extension field, which will involve computing with high-degree polynomials with large integer coefficients.

Writing efficient software for arithmetic as it is used typically in CAS requires the knowledge of many algorithms with their range of applicability, good programming skills in order to spend time only where it should be spent, and finally good knowledge of the target hardware. Indeed, it makes little sense to disregard the specifics of the intended hardware platforms, even more so since in the past years, we have seen a paradigm shift in terms of available hardware: so far, it used to be reasonable to consider that an end-user running a CAS would have access to a single-CPU processor. Nowadays, even a basic laptop computer has a multi-core processor and a powerful graphics card, and a workstation with a reconfigurable coprocessor is no longer science-fiction.

In this context, one of our goals is to investigate and take advantage of these influences and interactions between various available computing resources in order to design better algorithms for basic arithmetic objects. Of course, this is not disconnected from the other goals, since they all rely more or less on integer or polynomial arithmetic.

CARTE Project-Team

3. Research Program

3.1. Computer Virology

From a historical point of view, the first official virus appeared in 1983 on Vax-PDP 11. At the same time, a series of papers was published which always remains a reference in computer virology: Thompson [71], Cohen [39] and Adleman [28]. The literature which explains and discusses practical issues is quite extensive [44], [46]. However, there are only a few theoretical/scientific studies, which attempt to give a model of computer viruses.

A virus is essentially a self-replicating program inside an adversary environment. Self-replication has a solid background based on works on fixed point in λ-calculus and on studies of von Neumann [75]. More precisely we establish in [35] that Kleene's second recursion theorem [59] is the cornerstone from which viruses and infection scenarios can be defined and classified. The bottom line of a virus behavior is

1. a virus infects programs by modifying them,
2. a virus copies itself and can mutate,
3. it spreads throughout a system.

The above scientific foundation justifies our position to use the word virus as a generic word for self-replicating malwares. There is yet a difference. A malware has a payload, and virus may not have one. For example, a worm is an autonous self-replicating malware and so falls into our definition. In fact, the current malware taxonomy (virus, worms, trojans, ...) is unclear and subject to debate.

3.2. Computation over continuous structures

Classical recursion theory deals with computability over discrete structures (natural numbers, finite symbolic words). There is a growing community of researchers working on the extension of this theory to continuous structures arising in mathematics. One goal is to give foundations of numerical analysis, by studying the limitations of machines in terms of computability or complexity, when computing with real numbers. Classical questions are : if a function $f : \mathbb{R} \to \mathbb{R}$ is computable in some sense, are its roots computable? in which time? Another goal is to investigate the possibility of designing new computation paradigms, transcending the usual discrete-time, discrete-space computer model initiated by the Turing machine that is at the base of modern computers.

While the notion of a computable function over discrete data is captured by the model of Turing machines, the situation is more delicate when the data are continuous, and several non-equivalent models exist. In this case, let us mention computable analysis, which relates computability to topology [43], [74]; the Blum-Shub-Smale model (BSS), where the real numbers are treated as elementary entities [34]; the General Purpose Analog Computer (GPAC) introduced by Shannon [69]with continuous time.

3.3. Rewriting

The rewriting paradigm is now widely used for specifying, modelizing, programming and proving. It allows one to easily express deduction systems in a declarative way, and to express complex relations on infinite sets of states in a finite way, provided they are countable. Programming languages and environments with a rewriting based semantics have been developed ; see ASF+SDF [36], MAUDE [38], and TOM [66].

For basic rewriting, many techniques have been developed to prove properties of rewrite systems like confluence, completeness, consistency or various notions of termination. Proof methods have also been proposed for extensions of rewriting such as equational extensions, consisting of rewriting modulo a set of axioms, conditional extensions where rules are applied under certain conditions only, typed extensions, where rules are applied only if there is a type correspondence between the rule and the term to be rewritten, and constrained extensions, where rules are enriched by formulas to be satisfied [30], [42], [70].

An interesting aspect of the rewriting paradigm is that it allows automatable or semi-automatable correctness proofs for systems or programs: the properties of rewriting systems as those cited above are translatable to the deduction systems or programs they formalize and the proof techniques may directly apply to them.

Another interesting aspect is that it allows characteristics or properties of the modelled systems to be expressed as equational theorems, often automatically provable using the rewriting mechanism itself or induction techniques based on completion [41]. Note that the rewriting and the completion mechanisms also enable transformation and simplification of formal systems or programs.

Applications of rewriting-based proofs to computer security are various. Approaches using rule-based specifications have recently been proposed for detection of computer viruses [72], [73]. For several years, in our team, we have also been working in this direction. We already proposed an approach using rewriting techniques to abstract program behaviors for detecting suspicious or malicious programs [31], [32].

CASCADE Project-Team

3. Research Program

3.1. Randomness in Cryptography

Randomness is a key ingredient for cryptography. Random bits are necessary not only for generating cryptographic keys, but are also often an part of steps of cryptographic algorithms. In some cases, probabilistic protocols make it possible to perform tasks that are impossible deterministically. In other cases, probabilistic algorithms are faster, more space efficient or simpler than known deterministic algorithms. Cryptographers usually assume that parties have access to perfect randomness but in practice this assumption is often violated and a large body of research is concerned with obtaining such a sequence of random or pseudorandom bits.

One of the project-team research goals is to get a better understanding of the interplay between randomness and cryptography and to study the security of various cryptographic protocols at different levels (information-theoretic and computational security, number-theoretic assumptions, design and provable security of new and existing constructions).

Cryptographic literature usually pays no attention to the fact that in practice randomness is quite difficult to generate and that it should be considered as a resource like space and time. Moreover since the perfect randomness abstraction is not physically realizable, it is interesting to determine whether imperfect randomness is "good enough" for certain cryptographic algorithms and to design algorithms that are robust with respect to deviations of the random sources from true randomness.

The power of randomness in computation is a central problem in complexity theory and in cryptography. Cryptographers should definitely take these considerations into account when proposing new cryptographic schemes: there exist computational tasks that we only know how to perform efficiently using randomness but conversely it is sometimes possible to remove randomness from probabilistic algorithms to obtain efficient deterministic counterparts. Since these constructions may hinder the security of cryptographic schemes, it is of high interest to study the efficiency/security tradeoff provided by randomness in cryptography.

Quite often in practice, the random bits in cryptographic protocols are generated by a pseudorandom number generation process. When this is done, the security of the scheme of course depends in a crucial way on the quality of the random bits produced by the generator. Despite the importance, many protocols used in practice often leave unspecified what pseudorandom number generation to use. It is well-known that pseudorandom generators exist if and only if one-way functions exist and there exist efficient constructions based on various number-theoretic assumptions. Unfortunately, these constructions are too inefficient and many protocols used in practice rely on "ad-hoc" constructions. It is therefore interesting to propose more efficient constructions, to analyze the security of existing ones and of specific cryptographic constructions that use weak pseudorandom number generators.

The project-team undertakes research in these three aspects. The approach adopted is both theoretical and practical, since we provide security results in a mathematical frameworks (information theoretic or computational) with the aim to design protocols among the most efficient known.

3.2. Lattice Cryptography

The security of almost all public-key cryptographic protocols in use today relies on the presumed hardness of problems from number theory such as factoring and discrete log. This is somewhat problematic because these problems have very similar underlying structure, and its unforeseen exploit can render all currently used public key cryptography insecure. This structure was in fact exploited by Shor to construct efficient quantum algorithms that break all hardness assumptions from number theory that are currently in use. And so naturally, an important area of research is to build provably-secure protocols based on mathematical problems that are unrelated to factoring and discrete log. One of the most promising directions in this line of research is using lattice problems as a source of computational hardness —in particular since they also offer features that other alternative public-key cryptosystems (such as MQ-based, code-based or hash-based schemes) cannot provide.

At its very core, secure communication rests on two foundations: authenticity and secrecy. Authenticity assures the communicating parties that they are indeed communicating with each other and not with some potentially malicious outside party. Secrecy is necessary so that no one except the intended recipient of a message is able to deduce anything about its contents.

Lattice cryptography might find applications towards constructing practical schemes for resolving essential cryptographic problems —in particular, guaranteeing authenticity. On this front, our team is actively involved in pursuing the following two objectives:

1. Construct, implement, and standardize a practical public key digital signature scheme that is secure against quantum adversaries.
2. Construct, implement, and standardize a symmetric key authentication scheme that is secure against side channel attacks and is more efficient than the basic scheme using AES with masking.

Despite the great progress in constructing fairly practical lattice-based encryption and signature schemes, efficiency still remains a very large obstacle for advanced lattice primitives. While constructions of identity-based encryption schemes, group signature schemes, functional encryption schemes, and even fully-homomorphic encryption schemes are known, the implementations of these schemes are extremely inefficient.

Fully Homomorphic Encryption (FHE) is a very active research area. Let us just give one example illustrating the usefulness of computing on encrypted data: Consider an on-line patent database on which firms perform complex novelty queries before filing patents. With current technologies, the database owner might analyze the queries, infer the invention and apply for a patent before the genuine inventor. While such frauds were not reported so far, similar incidents happen during domain name registration. Several websites propose "registration services" preceded by "availability searches". These queries trigger the automated registration of the searched domain names which are then proposed for sale. Algorithms allowing arbitrary computations without disclosing their inputs (and/or their results) are hence of immediate usefulness.

In 2009, IBM announced the discovery of a FHE scheme by Craig Gentry. The security of this algorithm relies on worst-case problems over ideal lattices and on the hardness of the sparse subset sum problem. Gentry's construction is an ingenious combination of two ideas: a somewhat homomorphic scheme (capable of supporting many "logical or" operations but very few "ands") and a procedure that refreshes the homomorphically processed ciphertexts. Gentry's main conceptual achievement is a "bootstrapping" process in which the somewhat homomorphic scheme evaluates its own decryption circuit (self-reference) to refresh (recrypt) ciphertexts.

Unfortunately, it is safe to surmise that if the state of affairs remains as it is in the present, then despite all the theoretical efforts that went into their constructions, these schemes will never be used in practical applications.

Our team is looking at the foundations of these primitives with the hope of achieving a breakthrough that will allow them to be practical in the near future.

3.3. Security amidst Concurrency on the Internet

Cryptographic protocols that are secure when executed in isolation, can be completely insecure when multiple such instances are executed concurrently (as is unavoidable on the Internet) or when used as a part of a larger protocol. For instance, a man-in-the-middle attacker participating in two simultaneous executions of a cryptographic protocol might use messages from one of the executions in order to compromise the security of the second – Lowe's attack on the Needham-Schroeder authentication protocol and Bleichenbacher's attack on SSL work this way. Our research addresses security amidst concurrent executions in secure computation and key exchange protocols.

Secure computation allows several mutually distrustful parties to collaboratively compute a public function of their inputs, while providing the same security guarantees as if a trusted party had performed the computation. Potential applications for secure computation include anonymous voting as well as privacy-preserving auctions and data-mining. Our recent contributions on this topic include

1. new protocols for secure computation in a model where each party interacts only once, with a single centralized server; this model captures communication patterns that arise in many practical settings, such as that of Internet users on a website,

2. and efficient constructions of universally composable commitments and oblivious transfer protocols, which are the main building blocks for general secure computation.

In key exchange protocols, we are actively involved in designing new password-authenticated key exchange protocols, as well as the analysis of the widely-used SSL/TLS protocols.

3.4. Symmetric Key Cryptanalysis

Symmetric key cryptographic primitives play a very important role in secure communications. For example, block ciphers and stream ciphers are used to protect the privacy of cellular phone users from eavesdroppers, while MACs (message authentication codes) ensure that active attackers cannot interfere with cellular communication without being detected.

Since there is no method of formally proving that a complex modern symmetric key cipher is secure, there is no choice but to consider it secure if there are no known attacks against it. Thus, a symmetric key cipher should undergo an extensive cryptanalytic effort to evaluate its resistance against both well-known and new types of attacks. The goal of cryptanalytic is thus to ensure that only the strongest symmetric key cryptographic primitives are deployed and used in practice.

The team contributes to this field by proposing new cryptanalytic techniques and applying them to both new and existing secret key primitives, helping to understand their security.

CASSIS Project-Team

3. Research Program

3.1. Introduction

Our main goal is to design techniques and to develop tools for the verification of (safety-critical) systems, such as programs or protocols. To this end, we develop a combination of techniques based on automated deduction for program verification, constraint resolution for test generation, and reachability analysis for the verification of infinite-state systems.

3.2. Automated Deduction

The main goal is to prove the validity of assertions obtained from program analysis. To this end, we develop techniques and automated deduction systems based on rewriting and constraint solving. The verification of recursive data structures relies on inductive reasoning or the manipulation of equations and it also exploits some form of reasoning modulo properties of selected operators (such as associativity and/or commutativity).

Rewriting, which allows us to simplify expressions and formulae, is a key ingredient for the effectiveness of many state-of-the-art automated reasoning systems. Furthermore, a well-founded rewriting relation can be also exploited to implement reasoning by induction. This observation forms the basis of our approach to inductive reasoning, with high degree of automation and the possibility to refute false conjectures.

The constraints are the key ingredient to postpone the activity of solving complex symbolic problems until it is really necessary. They also allow us to increase the expressivity of the specification language and to refine theorem-proving strategies. As an example of this, the handling of constraints for unification problems or for the orientation of equalities in the presence of interpreted operators (e.g., commutativity and/or associativity function symbols) will possibly yield shorter automated proofs.

Finally, decision procedures are being considered as a key ingredient for the successful application of automated reasoning systems to verification problems. A decision procedure is an algorithm capable of efficiently deciding whether formulae from certain theories (such as Presburger arithmetic, lists, arrays, and their combination) are valid or not. We develop techniques to build and to combine decision procedures for the domains which are relevant to verification problems. We also perform experimental evaluation of the proposed techniques by combining propositional reasoning (implemented by means of Boolean solvers, e.g., SAT solvers) and decision procedures to get solvers for the problem of Satisfiability Modulo Theories (SMT).

3.3. Synthesizing and Solving Constraints

Applying constraint logic programming technology in the validation and verification area is currently an active way of research. It usually requires the design of specific solvers to deal with the description language's vocabulary. For instance, we are interested in applying a solver for set constraintsto evaluate set-oriented formal specifications. By evaluation, we mean the encoding of the formal model into a constraint system, and the ability for the solver to verify the invariant on the current constraint graph, to propagate preconditions or guards, and to apply a substitution calculus on this graph. The constraint solver is used for animating specifications and automatically generating abstract test cases.

3.4. Rewriting-based Safety Checking

Invariant checking and strenghtening is the dual of reachability analysis, and can thus be used for verifying safety properties of infinite-state systems. In fact, many infinite-state systems are just parameterized systems which become finite state systems when parameters are instantiated. Then, the challenge is to automatically discharge the maximal number of proof obligations coming from the decomposition of the invariance conditions. For parameterized systems, we are interested in a deductive approach where states are defined by first order formulae with equality, and proof obligations are checked by SMT solvers.

CELTIQUE Project-Team

3. Research Program

3.1. Static program analysis

Static program analysis is concerned with obtaining information about the run-time behaviour of a program without actually running it. This information may concern the values of variables, the relations among them, dependencies between program values, the memory structure being built and manipulated, the flow of control, and, for concurrent programs, synchronisation among processes executing in parallel. Fully automated analyses usually render approximate information about the actual program behaviour. The analysis is correct if the information includes all possible behaviour of a program. Precision of an analysis is improved by reducing the amount of information describing spurious behaviour that will never occur.

Static analysis has traditionally found most of its applications in the area of program optimisation where information about the run-time behaviour can be used to transform a program so that it performs a calculation faster and/or makes better use of the available memory resources. The last decade has witnessed an increasing use of static analysis in software verification for proving invariants about programs. The Celtique project is mainly concerned with this latter use. Examples of static analysis include:

- Data-flow analysis as it is used in optimising compilers for imperative languages. The properties can either be approximations of the values of an expression ("the value of variable x is greater than 0" or x is equal to y at this point in the program") or more intensional information about program behaviour such as "this variable is not used before being re-defined" in the classical "dead-variable" analysis [72].
- Analyses of the memory structure includes shape analysis that aims at approximating the data structures created by a program. Alias analysis is another data flow analysis that finds out which variables in a program addresses the same memory location. Alias analysis is a fundamental analysis for all kinds of programs (imperative, object-oriented) that manipulate state, because alias information is necessary for the precise modelling of assignments.
- Control flow analysis will find a safe approximation to the order in which the instructions of a program are executed. This is particularly relevant in languages where parameters or functions can be passed as arguments to other functions, making it impossible to determine the flow of control from the program syntax alone. The same phenomenon occurs in object-oriented languages where it is the class of an object (rather than the static type of the variable containing the object) that determines which method a given method invocation will call. Control flow analysis is an example of an analysis whose information in itself does not lead to dramatic optimisations (although it might enable in-lining of code) but is necessary for subsequent analyses to give precise results.

Static analysis possesses strong **semantic foundations**, notably abstract interpretation [54], that allow to prove its correctness. The implementation of static analyses is usually based on well-understood constraint-solving techniques and iterative fixpoint algorithms. In spite of the nice mathematical theory of program analysis and the solid algorithmic techniques available one problematic issue persists, *viz.*, the *gap* between the analysis that is proved correct on paper and the analyser that actually runs on the machine. While this gap might be small for toy languages, it becomes important when it comes to real-life languages for which the implementation and maintenance of program analysis tools become a software engineering task. A *certified static analysis* is an analysis that has been formally proved correct using a proof assistant.

In previous work we studied the benefit of using abstract interpretation for developing **certified static analyses** [52], [75]. The development of certified static analysers is an ongoing activity that will be part of the Celtique project. We use the Coq proof assistant which allows for extracting the computational content of a constructive proof. A Caml implementation can hence be extracted from a proof of existence, for any program, of a correct approximation of the concrete program semantics. We have isolated a theoretical framework based on abstract interpretation allowing for the formal development of a broad range of static analyses. Several case studies for the analysis of Java byte code have been presented, notably a memory usage analysis [53]. This work has recently found application in the context of Proof Carrying Code and have also been successfully applied to particular form of static analysis based on term rewriting and tree automata [5].

3.1.1. Static analysis of Java

Precise context-sensitive control-flow analysis is a fundamental prerequisite for precisely analysing Java programs. Bacon and Sweeney's Rapid Type Analysis (RTA) [45] is a scalable algorithm for constructing an initial call-graph of the program. Tip and Palsberg [80] have proposed a variety of more precise but scalable call graph construction algorithms *e.g.,* MTA, FTA, XTA which accuracy is between RTA and 0'CFA. All those analyses are not context-sensitive. As early as 1991, Palsberg and Schwartzbach [73], [74] proposed a theoretical parametric framework for typing object-oriented programs in a context-sensitive way. In their setting, context-sensitivity is obtained by explicit code duplication and typing amounts to analysing the expanded code in a context-insensitive manner. The framework accommodates for both call-contexts and allocation-contexts.

To assess the respective merits of different instantiations, scalable implementations are needed. For Cecil and Java programs, Grove *et al.,* [61], [60] have explored the algorithmic design space of contexts for benchmarks of significant size. Latter on, Milanova *et. al.,* [67] have evaluated, for Java programs, a notion of context called *object-sensitivity* which abstracts the call-context by the abstraction of the `this` pointer. More recently, Lhotak and Hendren [65] have extended the empiric evaluation of object-sensitivity using a BDD implementation allowing to cope with benchmarks otherwise out-of-scope. Besson and Jensen [49] proposed to use DATALOG in order to specify context-sensitive analyses. Whaley and Lam [81] have implemented a context-sensitive analysis using a BDD-based DATALOG implementation.

Control-flow analyses are a prerequisite for other analyses. For instance, the security analyses of Livshits and Lam [66] and the race analysis of Naik, Aiken [68] and Whaley [69] both heavily rely on the precision of a control-flow analysis.

Control-flow analysis allows to statically prove the absence of certain run-time errors such as "message not understood" or cast exceptions. Yet it does not tackle the problem of "null pointers". Fahnrich and Leino [57] propose a type-system for checking that after object creation fields are non-null. Hubert, Jensen and Pichardie have formalised the type-system and derived a type-inference algorithm computing the most precise typing [64]. The proposed technique has been implemented in a tool called NIT [63]. Null pointer detection is also done by bug-detection tools such as FindBugs [63]. The main difference is that the approach of findbugs is neither sound nor complete but effective in practice.

3.1.2. Quantitative aspects of static analysis

Static analyses yield qualitative results, in the sense that they compute a safe over-approximation of the concrete semantics of a program, w.r.t. an order provided by the abstract domain structure. Quantitative aspects of static analysis are two-sided: on one hand, one may want to express and verify (compute) quantitative properties of programs that are not captured by usual semantics, such as time, memory, or energy consumption; on the other hand, there is a deep interest in quantifying the precision of an analysis, in order to tune the balance between complexity of the analysis and accuracy of its result.

The term of quantitative analysis is often related to probabilistic models for abstract computation devices such as timed automata or process algebras. In the field of programming languages which is more specifically addressed by the Celtique project, several approaches have been proposed for quantifying resource usage: a non-exhaustive list includes memory usage analysis based on specific type systems [62], [44], linear

logic approaches to implicit computational complexity [46], cost model for Java byte code [40] based on size relation inference, and WCET computation by abstract interpretation based loop bound interval analysis techniques [55].

We have proposed an original approach for designing static analyses computing program costs: inspired from a probabilistic approach [76], a quantitative operational semantics for expressing the cost of execution of a program has been defined. Semantics is seen as a linear operator over a dioid structure similar to a vector space. The notion of long-run cost is particularly interesting in the context of embedded software, since it provides an approximation of the asymptotic behaviour of a program in terms of computation cost. As for classical static analysis, an abstraction mechanism allows to effectively compute an over-approximation of the semntics, both in terms of costs and of accessible states [51]. An example of cache miss analysis has been developed within this framework [79].

3.2. Software certification

The term "software certification" has a number of meanings ranging from the formal proof of program correctness via industrial certification criteria to the certification of software developers themselves! We are interested in two aspects of software certification:

- industrial, mainly process-oriented certification procedures
- software certificates that convey semantic information about a program

Semantic analysis plays a role in both varieties.

Criteria for software certification such as the Common criteria or the DOA aviation industry norms describe procedures to be followed when developing and validating a piece of software. The higher levels of the Common Criteria require a semi-formal model of the software that can be refined into executable code by traceable refinement steps. The validation of the final product is done through testing, respecting criteria of coverage that must be justified with respect to the model. The use of static analysis and proofs has so far been restricted to the top level 7 of the CC and has not been integrated into the aviation norms.

3.2.1. Process-oriented software certification

The testing requirements present in existing certification procedures pose a challenge in terms of the automation of the test data generation process for satisfying functional and structural testing requirements. For example, the standard document which currently governs the development and verification process of software in airborne system (DO-178B) requires the coverage of all the statements, all the decisions of the program at its higher levels of criticality and it is well-known that DO-178B structural coverage is a primary cost driver on avionics project. Although they are widely used, existing marketed testing tools are currently restricted to test coverage monitoring and measurements [0] but none of these tools tries to find the test data that can execute a given statement, branch or path in the source code. In most industrial projects, the generation of structural test data is still performed manually and finding automatic methods for this problem remains a challenge for the test community. Building automatic test case generation methods requires the development of precise semantic analysis which have to scale up to software that contains thousands of lines of code.

Static analysis tools are so far not a part of the approved certification procedures. For this to change, the analysers themselves must be accepted by the certification bodies in a process called “Qualification of the tools” in which the tools are shown to be as robust as the software it will certify. We believe that proof assistants have a role to play in building such certified static analysis as we have already shown by extracting provably correct analysers for Java byte code.

[0] Coverage monitoring answers to the question: what are the statements or branches covered by the test suite ? While coverage measurements answers to: how many statements or branches have been covered ?

3.2.2. Semantic software certificates

The particular branch of information security called "language-based security" is concerned with the study of programming language features for ensuring the security of software. Programming languages such as Java offer a variety of language constructs for securing an application. Verifying that these constructs have been used properly to ensure a given security property is a challenge for program analysis. One such problem is confidentiality of the private data manipulated by a program and a large group of researchers have addressed the problem of tracking information flow in a program in order to ensure that *e.g.*, a credit card number does not end up being accessible to all applications running on a computer [78], [48]. Another kind of problems concern the way that computational resources are being accessed and used, in order to ensure that a given access policy is being implemented correctly and that a given application does not consume more resources that it has been allocated. Members of the Celtique team have proposed a verification technique that can check the proper use of resources of Java applications running on mobile telephones [50]. **Semantic software certificates** have been proposed as a means of dealing with the security problems caused by mobile code that is downloaded from foreign sites of varying trustworthiness and which can cause damage to the receiving host, either deliberately or inadvertently. These certificates should contain enough information about the behaviour of the downloaded code to allow the code consumer to decide whether it adheres to a given security policy.

Proof-Carrying Code (PCC) [70] is a technique to download mobile code on a host machine while ensuring that the code adheres to a specified security policy. The key idea is that the code producer sends the code along with a proof (in a suitably chosen logic) that the code is secure. Upon reception of the code and before executing it, the consumer submits the proof to a proof checker for the logic. Our project focus on two components of the PCC architecture: the proof checker and the proof generator.

In the basic PCC architecture, the only components that have to be trusted are the program logic, the proof checker of the logic, and the formalization of the security property in this logic. Neither the mobile code nor the proposed proof—and even less the tool that generated the proof—need be trusted.

In practice, the *proof checker* is a complex tool which relies on a complex Verification Condition Generator (VCG). VCGs for real programming languages and security policies are large and non-trivial programs. For example, the VCG of the Touchstone verifier represents several thousand lines of C code, and the authors observed that "there were errors in that code that escaped the thorough testing of the infrastructure" [71]. Many solutions have been proposed to reduce the size of the trusted computing base. In the *foundational proof carrying code* of Appel and Felty [43], [42], the code producer gives a direct proof that, in some "foundational" higher-order logic, the code respects a given security policy. Wildmoser and Nipkow [83], [82]. prove the soundness of a *weakest precondition* calculus for a reasonable subset of the Java bytecode. Necula and Schneck [71] extend a small trusted core VCG and describe the protocol that the untrusted verifier must follow in interactions with the trusted infrastructure.

One of the most prominent examples of software certificates and proof-carrying code is given by the Java byte code verifier based on *stack maps*. Originally proposed under the term “lightweight Byte Code Verification” by Rose [77], the techniques consists in providing enough typing information (the stack maps) to enable the byte code verifier to check a byte code in one linear scan, as opposed to inferring the type information by an iterative data flow analysis. The Java Specification Request 202 provides a formalization of how such a verification can be carried out.

Inspired by this, Albert *et al.* [41] have proposed to use static analysis (in the form of abstract interpretation) as a general tool in the setting of mobile code security for building a proof-carrying code architecture. In their *abstraction-carrying code* framework, a program comes equipped with a machine-verifiable certificate that proves to the code consumer that the downloaded code is well-behaved.

3.2.3. Certified static analysis

In spite of the nice mathematical theory of program analysis (notably abstract interpretation) and the solid algorithmic techniques available one problematic issue persists, *viz.*, the *gap* between the analysis that is proved correct on paper and the analyser that actually runs on the machine. While this gap might be small for

toy languages, it becomes important when it comes to real-life languages for which the implementation and maintenance of program analysis tools become a software engineering task.

A *certified static analysis* is an analysis whose implementation has been formally proved correct using a proof assistant. Such analysis can be developed in a proof assistant like Coq [39] by programming the analyser inside the assistant and formally proving its correctness. The Coq extraction mechanism then allows for extracting a Caml implementation of the analyser. The feasibility of this approach has been demonstrated in [7].

We also develop this technique through certified reachability analysis over term rewriting systems. Term rewriting systems are a very general, simple and convenient formal model for a large variety of computing systems. For instance, it is a very simple way to describe deduction systems, functions, parallel processes or state transition systems where rewriting models respectively deduction, evaluation, progression or transitions. Furthermore rewriting can model every combination of them (for instance two parallel processes running functional programs).

Depending on the computing system modelled using rewriting, reachability (and unreachability) permits to achieve some verifications on the system: respectively prove that a deduction is feasible, prove that a function call evaluates to a particular value, show that a process configuration may occur, or that a state is reachable from the initial state. As a consequence, reachability analysis has several applications in equational proofs used in the theorem provers or in the proof assistants as well as in verification where term rewriting systems can be used to model programs.

For proving unreachability, i.e. safety properties, we already have some results based on the over-approximation of the set of reachable terms [58], [59]. We defined a simple and efficient algorithm [56] for computing exactly the set of reachable terms, when it is regular, and construct an over-approximation otherwise. This algorithm consists of a *completion* of a *tree automaton*, taking advantage of the ability of tree automata to finitely represent infinite sets of reachable terms.

To certify the corresponding analysis, we have defined a checker guaranteeing that a tree automaton is a valid fixpoint of the completion algorithm. This consists in showing that for all term recognised by a tree automaton all his rewrites are also recognised by the same tree automaton. This checker has been formally defined in Coq and an efficient Ocaml implementation has been automatically extracted [5]. This checker is now used to certify all analysis results produced by the regular completion tool as well as the optimised version of [47].

COMETE Project-Team

3. Research Program

3.1. Probability and information theory

Participants: Nicolas Bordenabe, Konstantinos Chatzikokolakis, Thomas Given-Wilson, Yusuke Kawamoto, Catuscia Palamidessi, Marco Stronati.

Much of the research of Comète focuses on security and privacy. In particular, we are interested in the problem of the leakage of secret information through public observables.

Ideally we would like systems to be completely secure, but in practice this goal is often impossible to achieve. Therefore, we need to reason about the amount of information leaked, and the utility that it can have for the adversary, i.e. the probability that the adversary is able to exploit such information.

The recent tendency is to use an information theoretic approach to model the problem and define the leakage in a quantitative way. The idea is to consider the system as an information-theoretic *channel*. The input represents the secret, the output represents the observable, and the correlation between the input and output (*mutual information*) represents the information leakage.

Information theory depends on the notion of entropy as a measure of uncertainty. From the security point of view, this measure corresponds to a particular model of attack and a particular way of estimating the security threat (vulnerability of the secret). Most of the proposals in the literature use Shannon entropy, which is the most established notion of entropy in information theory. We, however, consider also other notions, in particular Rényi min-entropy, which seems to be more appropriate for security in common scenarios like one-try attacks.

3.2. Expressiveness of Concurrent Formalisms

Participants: Catuscia Palamidessi, Luis Pino, Frank Valencia.

We study computational models and languages for distributed, probabilistic and mobile systems, with a particular attention to expressiveness issues. We aim at developing criteria to assess the expressive power of a model or formalism in a distributed setting, to compare existing models and formalisms, and to define new ones according to an intended level of expressiveness, also taking into account the issue of (efficient) implementability.

3.3. Concurrent constraint programming

Participants: Michell Guzman, Yamil Salim Perchy, Luis Pino, Frank Valencia.

Concurrent constraint programming (ccp) is a well established process calculus for modeling systems where agents interact by posting and asking information in a store, much like in users interact in *social networks*. This information is represented as first-order logic formulae, called constraints, on the shared variables of the system (e.g., $X > 42$). The most distinctive and appealing feature of ccp is perhaps that it unifies in a single formalism the operational view of processes based upon process calculi with a declarative one based upon first-order logic. It also has an elegant denotational semantics that interprets processes as closure operators (over the set of constraints ordered by entailment). In other words, any ccp process can be seen as an idempotent, increasing, and monotonic function from stores to stores. Consequently, ccp processes can be viewed as: computing agents, formulae in the underlying logic, and closure operators. This allows ccp to benefit from the large body of techniques of process calculi, logic and domain theory.

Our research in ccp develops along the following two lines:

1. **(a)** The study of a bisimulation semantics for ccp. The advantage of bisimulation, over other kinds of semantics, is that it can be efficiently verified.
2. **(b)** The extension of ccp with constructs to capture emergent systems such as those in social networks and cloud computing.

3.4. Model checking

Participants: Konstantinos Chatzikokolakis, Catuscia Palamidessi.

Model checking addresses the problem of establishing whether a given specification satisfies a certain property. We are interested in developing model-checking techniques for verifying concurrent systems of the kind explained above. In particular, we focus on security and privacy, i.e., on the problem of proving that a given system satisfies the intended security or privacy properties. Since the properties we are interested in have a probabilistic nature, we use probabilistic automata to model the protocols. A challenging problem is represented by the fact that the interplay between nondeterminism and probability, which in security presents subtleties that cannot be handled with the traditional notion of a scheduler,

COMPSYS Project-Team

3. Research Program

3.1. Architecture and compilation trends

The embedded system design community is facing two challenges:

- The complexity of embedded applications is increasing at a rapid rate.
- The needed increase in processing power is no longer obtained by increases in the clock frequency, but by increased parallelism.

While, in the past, each type of embedded application was implemented in a separate appliance, the present tendency is toward a universal hand-held object, which must serve as a cell-phone, as a personal digital assistant, as a game console, as a camera, as a Web access point, and much more. One may say that embedded applications are of the same level of complexity as those running on a PC, but they must use a more constrained platform in terms of processing power, memory size, and energy consumption. Furthermore, most of them depend on international standards (e.g., in the field of radio digital communication), which are evolving rapidly. Lastly, since ease of use is at a premium for portable devices, these applications must be integrated seamlessly to a degree that is unheard of in standard computers.

All of this dictates that modern embedded systems retain some form of programmability. For increased designer productivity and reduced time-to-market, programming must be done in some high-level language, with appropriate tools for compilation, run-time support, and debugging. This does not mean however that all embedded systems (or all of an embedded system) must be processor based. Another solution is the use of field programmable gate arrays (FPGA), which may be programmed at a much finer grain than a processor, although the process of FPGA "programming" is less well understood than software generation. Processors are better than application-specific circuits at handling complicated control and unexpected events. On the other hand, FPGAs may be tailored to just meet the needs of their application, resulting in better energy and silicon area usage. It is expected that most embedded systems will use a combination of general-purpose processors, specific processors like DSPs, and FPGA accelerators (or even low-power GPUs). Such a combination DSP+FPGA is already present in recent versions of the Atom Intel processor.

As a consequence, parallel programming, which has long been confined to the high-performance community, must become the common place rather than the exception. In the same way that sequential programming moved from assembly code to high-level languages at the price of a slight loss in performance, parallel programming must move from low-level tools, like OpenMP or even MPI, to higher-level programming environments. While fully-automatic parallelization is a Holy Grail that will probably never be reached in our lifetimes, it will remain as a component in a comprehensive environment, including general-purpose parallel programming languages, domain-specific parallelizers, parallel libraries and run-time systems, back-end compilation, dynamic parallelization. The landscape of embedded systems is indeed very diverse and many design flows and code optimization techniques must be considered. For example, embedded processors (micro-controllers, DSP, VLIW) require powerful back-end optimizations that can take into account hardware specificities, such as special instructions and particular organizations of registers and memories. FPGA and hardware accelerators, to be used as small components in a larger embedded platform, require "hardware compilation", i.e., design flows and code generation mechanisms to generate non-programmable circuits. For the design of a complete system-on-chip platform, architecture models, simulators, debuggers are required. The same is true for multicores of any kind, GPGPU ("general-purpose" graphical processing units), CGRA (coarse-grain reconfigurable architectures), which require specific methodologies and optimizations, although all these techniques converge or have connections. In other words, embedded systems need all usual aspects of the process that transforms some specification down to an executable, software or hardware. In this wide range of topics, Compsys concentrates on the code optimizations aspects (and the associated analysis) in this transformation chain, restricting to compilation (transforming a program to a program) for embedded

processors and programmable accelerators, and to high-level synthesis (transforming a program into a circuit description) for FPGAs.

Actually, it is not a surprise to see compilation and high-level synthesis getting closer (in the last 10 years now). Now that high-level synthesis has grown up sufficiently to be able to rely on place-and-route tools, or even to synthesize C-like languages, standard techniques for back-end code generation (register allocation, instruction selection, instruction scheduling, software pipelining) are used in HLS tools. At the higher level, programming languages for programmable parallel platforms share many aspects with high-level specification languages for HLS, for example, the description and manipulations of nested loops, or the model of computation/communication (e.g., Kahn process networks and its many "streaming" variants). In all aspects, the frontier between software and hardware is vanishing. For example, in terms of architecture, customized processors (with processor extension as first proposed by Tensilica) share features with both general-purpose processors and hardware accelerators. FPGAs are both hardware and software as they are fed with "programs" representing their hardware configurations.

In other words, this convergence in code optimizations explains why Compsys studies both program compilation and high-level synthesis, and at both front-end and back-end levels, the first one acting more at the granularity of memories, transfers, and multiple cores, the second one more at the granularity of registers, system calls, and single core. Both levels must be considered as they interact with each other. Front-end optimizations must be aware of what back-end optimizations will do, as single core performance remain the basis for good parallel performances. Some front-end optimizations even act directly on back-end features, for example register tiling considered as a source-level transformation. Also, from a conceptual point of view, the polyhedral techniques developed by Compsys are actually the symbolic front-end counterpart, for structured loops, of back-end analysis and optimizations of unstructured programs (through control-flow graphs), such as dependence analysis, scheduling, lifetime analysis, register allocation, etc. A strength of Compsys so far was to juggle with both aspects, one more on graph theory with SSA-type optimizations, the other with polyhedra representing loops, and to exploit the correspondence between both. This has still to be exploited, for applying polyhedral techniques to more irregular programs.

Besides, Compsys has a tradition of building free software tools for linear programming and optimization in general, and will continue it, as needed for our current research.

3.1.1. Compilation and languages issues in the context of embedded processors, "embedded systems", and programmable accelerators

Compilation is an old activity, in particular back-end code optimizations. The development of embedded systems was one of the reasons for the revival of compilation activities as a research topic. Applications for embedded computing systems generate complex programs and need more and more processing power. This evolution is driven, among others, by the increasing impact of digital television, the first instances of UMTS networks, and the increasing size of digital supports, like recordable DVD, and even Internet applications. Furthermore, standards are evolving very rapidly (see for instance the successive versions of MPEG). As a consequence, the industry has focused on programmable structures, whose flexibility more than compensates for their larger size and power consumption. The appliance provider has a choice between hard-wired structures (Asic), special-purpose processors (Asip), (quasi) general-purpose processors (DSP for multimedia applications), and now hardware accelerators (dedicated platforms – such as those developed by Thales or the CEA –, or more general-purpose accelerators such as GPUs or even multicores, even if these are closer to small HPC platforms than truly embedded systems). Our cooperation with STMicroelectronics, until 2012, focused on investigating the compilation for specialized processors, such as the ST100 (DSP processor) and the ST200 (VLIW DSP processor) family. Even for this restricted class of processors, the diversity is large, and the potential for instruction level parallelism (SIMD, MMX), the limited number of registers and the small size of the memory, the use of direct-mapped instruction caches, of predication, generate many open problems. Our goal was to contribute to their understanding and their solutions.

An important concept to cope with the diversity of platforms is the concept of *virtualization*, which is a key for more portability, more simplicity, more reliability, and of course more security. This concept – implemented at

low level through binary translation and just-in-time (JIT) compilation [0] – consists in hiding the architecture-dependent features as long as possible during the compilation process. It has been used for a while for servers such as HotSpot, a bit more recently for workstations, and now for embedded computing. The same needs drive the development of intermediate languages such as OpenCL to, not necessarily hide, but at least make more uniform, the different facets of the underlying architectures. The challenge is then to design and compile high-productivity and high-performance languages [0] (coping with parallelism and heterogeneity) that can be ported to such intermediate languages, or to architecture-dependent runtime systems. The offloading of computation kernels, through source-to-source compilation, targeting back-end C dialects, has the same goals: to automate application porting to the variety of accelerators.

For JIT compilation, the compactness of the information representation, and thus its pertinence, is an important criterion for such late compilation phases. Indeed, the intermediate representation (IR) is evolving not only from a target-independent description to a target-dependent one, but also from a situation where the compilation time is almost unlimited (cross-compilation) to one where any type of resource is limited. This is one of the reasons why static single assignment (SSA), a sparse compact representation of liveness information, became popular in embedded compilation. If time constraints are common to all JIT compilers (not only for embedded computing), the benefit of using SSA is also in terms of its good ratio pertinence/storage of information. It also enables to simplify algorithms, which is also important for increasing the reliability of the compiler. In this context, our aim has been, in particular, to develop exact or heuristic solutions to *combinatorial* problems that arise in compilation for VLIW and DSP processors, and to integrate these methods into industrial compilers for DSP processors (mainly ST100, ST200, Strong ARM). Such combinatorial problems can be found in register allocation, opcode selection, code placement, when removing the SSA multiplexer functions (known as ϕ functions). These optimizations are usually done in the last phases of the compiler, using an assembly-level intermediate representation. As mentioned in Sections 2.3 and 2.4 , we made a lot of progress in this area in our past collaborations with STMicroelectronics (see also previous activity reports). Through the Sceptre and Mediacom projects, we first revisited, in the light of SSA, some code optimizations in an aggressive context, to develop better strategies, without eliminating too quickly solutions that may have been considered as too expensive in the past. Then we exploited the new concepts introduced in the aggressive context to design better algorithms in a JIT context, focusing on the speed of algorithms and their memory footprint, without compromising too much on the quality of the generated code.

Our research directions are currently more focused on programmable accelerators, such as GPU and multi-cores, but still considering *static* compilation and without forgetting the link between high-level (in general at source-code level) and low-level (i.e., at assembly-code level) optimizations. They concern program analysis (of both sequential and parallel specifications), program optimizations (for memory hierarchies, parallelism, streaming, etc.), and also the link with applications and between compilers and users (programmers). Polyhedral techniques play an important role in these directions, even if control-flow-based techniques remain in the background and may come back at any time in the foreground. This is also the case for high-level synthesis, as exposed in the next section.

3.1.2. Context of high-level synthesis and FPGA platforms

High-level synthesis has become a necessity, mainly because the exponential increase in the number of gates per chip far outstrips the productivity of human designers. Besides, applications that need hardware accelerators usually belong to domains, like telecommunications and game platforms, where fast turn-around

[0] *Aggressive compilation* consists in allowing more time to implement more complete and costly solutions: the compiled program is loaded in permanent memory (ROM, flash, etc.) and its compilation time is less relevant than the execution time, size, and energy consumption of the produced code, which can have a critical impact on the cost and quality of the final product. Hence, the application is cross-compiled, i.e., compiled on a powerful platform distinct from the target processor. *Just-in-time compilation*, on the other hand, corresponds to compiling applets on demand on the target processor. For compatibility and compactness, the source languages are CIL or Java bytecode. The code can be uploaded or sold separately on a flash memory. Compilation is performed at load time and even dynamically during execution. The optimization heuristics, constrained by time and limited resources, are far from being aggressive. They must be fast but smart enough.

[0] For examples of such languages, see the keynotes event we organized in 2013: http://labexcompilation.ens-lyon.fr/hpc-languages.

and time-to-market minimization are paramount. When Compsys started, we were convinced that our expertise in compilation and automatic parallelization could contribute to the development of the needed tools.

Today, synthesis tools for FPGAs or ASICs come in many shapes. At the lowest level, there are proprietary Boolean, layout, and place-and-route tools, whose input is a VHDL or Verilog specification at the structural or register-transfer level (RTL). Direct use of these tools is difficult, for several reasons:

- A structural description is completely different from an usual algorithmic language description, as it is written in term of interconnected basic operators. One may say that it has a spatial orientation, in place of the familiar temporal orientation of algorithmic languages.
- The basic operators are extracted from a library, which poses problems of selection, similar to the instruction selection problem in ordinary compilation.
- Since there is no accepted standard for VHDL synthesis, each tool has its own idiosyncrasies and reports its results in a different format. This makes it difficult to build portable HLS tools.
- HLS tools have trouble handling loops. This is particularly true for logic synthesis systems, where loops are systematically unrolled (or considered as sequential) before synthesis. An efficient treatment of loops needs the polyhedral model. This is where past results from the automatic parallelization community are useful.
- More generally, a VHDL specification is too low level to allow the designer to perform, easily, higher-level code optimizations, especially on multi-dimensional loops and arrays, which are of paramount importance to exploit parallelism, pipelining, and perform communication and memory optimizations.

Some intermediate tools were proposed that generate VHDL from a specification in restricted C, both in academia (such as SPARK, Gaut, UGH, CloogVHDL), and in industry (such as C2H), CatapultC, Pico-Express, Vivado HLS. All these tools use only the most elementary form of parallelization, equivalent to instruction-level parallelism in ordinary compilers, with some limited form of block pipelining, and communication through FIFOs. Targeting one of these tools for low-level code generation, while we concentrate on exploiting loop parallelism, might be a more fruitful approach than directly generating VHDL. However, it may be that the restrictions they impose preclude efficient use of the underlying hardware. Our first experiments with these HLS tools reveal two important issues. First, they are, of course, limited to certain types of input programs so as to make their design flows successful, even if, over the years, they become more and more mature. But it remains a painful and tricky task for the user to transform the program so that it fits these constraints and to tune it to get good results. Automatic or semi-automatic program transformations can help the user achieve this task. Second, users, even expert users, have only a very limited understanding of what back-end compilers do and why they do not lead to the expected results. An effort must be done to analyze the different design flows of HLS tools, to explain what to expect from them, and how to use them to get a good quality of results. Our first goal is thus to develop high-level techniques that, used in front of existing HLS tools, improve their utilization. This should also give us directions on how to modify them or to design new tools from scratch.

More generally, we want to consider HLS as a more global parallelization process. So far, no HLS tools is capable of generating designs with communicating *parallel* accelerators, even if, in theory, at least for the scheduling part, a tool such as Pico-Express could have such capabilities. The reason is that it is, for example, very hard to automatically design parallel memories and to decide the distribution of array elements in memory banks to get the desired performances with parallel accesses. Also, how to express communicating processes at the language level? How to express constraints, pipeline behavior, communication media, etc.? To better exploit parallelism, a first solution is to extend the source language with parallel constructs, as in all derivations of the Kahn process networks model, including communicating regular processes (CRP, see later). The other solution is a form of automatic parallelization. However, classical methods, which are mostly based on scheduling, need to be revisited, to pay more attention to locality, process streaming, and low-level pipelining, which are of paramount importance in hardware. Besides, classical methods mostly rely on the runtime system to tailor the parallelism degree to the available resources. Obviously, there is no runtime system in hardware. The real challenge is thus to invent new scheduling algorithms that take resource, locality,

and pipelining into account, and then to infer the necessary hardware from the schedule. This is probably possible only for programs that fit into the polyhedral model, or in an incrementally-extended model.

Our research activities on polyhedral code analysis and optimizations directly target these HLS challenges. But they are not limited to the automatic generation of hardware as can be seen from our different contributions on X10, OpenStream, parametric tiling, etc. The same underlying concepts also arise when optimizing codes for GPUs and multicores. In this context of polyhedral analysis and optimizations, we will focus on three aspects:

- developing high-level transformations, especially for loops and memory/communication optimizations, that can be used in front of HLS tools so as to improve their use, as well as for hardware accelerators;
- developing concepts and techniques in a more global view of high-level synthesis and high-level parallel programming, starting from specification languages down to hardware implementation;
- developing more general code analysis so as to extract more information from codes as well as to extend the programs that can be handled.

3.2. Code analysis, code transformations, code optimizations

Embedded systems generated new problems in code analysis and optimization both for optimizing embedded software (compilation) and hardware (HLS). We now give a bit more details on some general challenges for program analysis, optimizations, and transformations, induced by this context, and on our methodology, in particular our development and use of polyhedral optimizations and its extensions.

3.2.1. Processes, scheduling, mapping, communications, etc.

Before mapping an application to an architecture, one has to decide which execution model is targeted and where to intervene in the design flow. Then one has to solve scheduling, placement, and memory management problems. These three aspects should be handled as a whole, but present state of the art dictates that they be treated separately. One of our aims will be to find more comprehensive solutions. The last task is code generation, both for the processing elements and the interfaces processors/accelerators.

There are basically two execution models for embedded systems: one is the classical accelerator model, in which data is deposited in the memory of the accelerator, which then does its job, and returns the results. In the streaming model, computations are done on the fly, as data items flow from an input channel to the output. Here, the data are never stored in (addressable) memory. Other models are special cases, or sometimes compositions of the basic models. For instance, a systolic array follows the streaming model, and sometimes extends it to higher dimensions. Software radio modems follow the streaming model in the large, and the accelerator model in detail. The use of first-in first-out queues (FIFO) in hardware design is an application of the streaming model. Experience shows that designs based on the streaming model are more efficient that those based on memory, for such applications. One of the point to be investigated is whether it is general enough to handle arbitrary (regular) programs. The answer is probably negative. One possible implementation of the streaming model is as a network of communicating processes either as Kahn process networks (FIFO based) or as our more recent model of communicating regular processes (memory based, see for example CRP below). It is an interesting fact that several researchers have investigated translation from process networks [22] and to process networks [27], [28]. Streaming languages such as StreamIt and OpenStream have also been developed.

Kahn process networks (KPN) were introduced 30 years ago as a notation for representing parallel programs. Such a network is built from processes that communicate via perfect FIFO channels. Because the channel histories are deterministic, one can define a semantics and talk meaningfully about the equivalence of two implementations. As a bonus, the dataflow diagrams used by signal processing specialists can be translated on-the-fly into process networks. The problem with KPNs is that they rely on an asynchronous execution model, while VLIW processors and FPGAs are synchronous or partially synchronous. Thus, there is a need for a tool for synchronizing KPNs. This can be done by computing a schedule that has to satisfy data dependences within each process, a causality condition for each channel (a message cannot be received before it is sent),

and real-time constraints. However, there is a difficulty in writing the channel constraints because one has to count messages in order to establish the send/receive correspondence and, in multi-dimensional loop nests, the counting functions may not be affine. Recent developments on the theory of polynomials (see Section 6.7) may offer a solution to this problem. One can also define another model, *communicating regular processes* (CRP), in which channels are represented as write-once/read-many arrays. One can then dispense with counting functions and prove that the determinacy property still holds. As an added benefit, a communication system in which the receive operation is not destructive is closer to the expectations of system designers.

The main difficulty with this approach is that ordinary programs are usually not constructed as process networks. One needs automatic or semi-automatic tools for converting sequential programs into process networks. One possibility is to start from array dataflow analysis [24] or variants. Another approach attempts to construct threads, i.e., pieces of sequential code with the smallest possible interactions. In favorable cases, one may even find outermost parallelism, i.e., threads with no interactions whatsoever. Tiling mechanisms can also be used to define atomic processes that can be pipelined as we proposed initially for FPGA [17].

Whatever the chosen solution (FIFO or addressable memory) for communicating between two accelerators or between the host processor and an accelerator, the problems of optimizing communication between processes and of optimizing buffers have to be addressed. Many local memory optimization problems have already been solved theoretically. Some examples are loop fusion and loop alignment for array contraction, techniques for data allocation in scratch-pad memory, or techniques for folding multi-dimensional arrays [21]. Nevertheless, the problem is still largely open. Some questions are: how to schedule a loop sequence (or even a process network) for minimal scratch-pad memory size? How is the problem modified when one introduces unlimited and/or bounded parallelism (same questions for analyzing explicitly-parallel programs)? How does one take into account latency or throughput constraints, bandwidth constraints for input and output channels, memory hierarchies? All loop transformations are useful in this context, in particular loop tiling, and may be applied either as source-to-source transformations (when used in front of HLS or C-level compilers) or to generate directly VHDL or lower-level C-dialects such as OpenCL. One should keep in mind that theory will not be sufficient to solve these problems. Experiments are required to check the relevance of the various models (computation model, memory model, power consumption model) and to select the most important factors according to the architecture. Besides, optimizations do interact: for instance, reducing memory size and increasing parallelism are often antagonistic. Experiments will be needed to find a global compromise between local optimizations. In particular, the design of cost models remain a fundamental challenge.

Finally, there remains the problem of code generation for accelerators. It is a well-known fact that modern methods for program optimization and parallelization do not generate a new program, but just deliver blueprints for program generation, in the form, e.g., of schedules, placement functions, or new array subscripting functions. A separate code generation phase must be crafted with care, as a too naive implementation may destroy the benefits of high-level optimization. There are two possibilities here as suggested before; one may target another high-level synthesis or compilation tool, or one may target directly VHDL or low-level code. Each approach has its advantages and drawbacks. However, both situations require that the input program respects some strong constraints on the code shape, array accesses, memory accesses, communication protocols, etc. Furthermore, to get the compilers do what the user wants requires a lot of program tuning, i.e., of program rewriting or of program annotations. What can be automated in this rewriting process? Semi-automated?

In other words, we still need to address scheduling, memory, communication, and code generation issues, in the light of the developments of new languages and architectures, pushing the limits of such an automation.

3.2.2. Beyond static control programs

With the advent of parallelism in supercomputers, the bulk of research in code transformation resulted in (semi-)automatic parallelization, with many techniques (analysis, scheduling, code generation, etc.) based on the description and manipulation of nested loops with polyhedra. Compsys has always taken an active part in the development of these so-called “polyhedral techniques”. Historically, these analysis were (wrongly) understood to be limited to static control programs.

Actually, the polyhedral model is neither a programming language nor an execution model rather an intermediate representation. As such, it can be generated from imperative sequential languages like C or Fortran, streaming languages like CRP, or equational languages like Alpha. While the structure of the model is the same in all three cases, it may enjoy different properties, e.g., a schedule always exists in the first case, not in the two others. The import of the polyhedral model is that many questions relative to the analysis of a program and the applicability of transformations can be answered precisely and efficiently by applying well-known mathematical results to the model.

For irregular programs, the basic idea is to construct a polyhedral over-approximation, i.e., a program which has more operations, a larger memory footprint, and more dependences than the original. One can then parallelize the approximated program using polyhedral tools, and then return to the original, either by introducing guards, or by insuring that approximations are harmless. This technique is the standard way of dealing with approximated dependences. We already started to study the impact of approximations in our kernel offloading technique, for optimizing remote communications [4]. It is clear however that this method will apply only to mildly non-polyhedral programs. The restriction to arrays as the only data structure is still present. Its advantage is that it will be able to subsume in a coherent framework many disparate tricks: the extraction of SCoPs, induction variable detection, the omission of non-affine subscripts, or the conversion of control dependences into data dependences. The link with the techniques developed in the PIPS compiler (based on array region analysis) is strong and will have to be explored.

Such over-approximations can be found by mean of abstract interpretation, a general framework to develop static analysis on real-life programs. However, they were designed mainly for verification purposes, thus precision was the main issue before scalability. Although many efforts were made in designing specialized analyses (pointers, data structures, arrays), these approaches still suffer from a lack of experimental evidence concerning their applicability for code optimization. Following our experience and work on termination analysis (that connects the work on back-end CFG-like and front-end polyhedral-like optimizations), and our work on range analysis of numerical variables and on the memory footprint on real-world C programs [9], our objective is to bridge the gap between abstract interpretation and compilation, by designing cheaper analyses that scale well, mainly based on compact representations derived from variants of static single assignment (SSA). We will focus on complex control, and complex data structures (pointers, lists) that still suffer from complexity issues in the area of optimisation.

Another possibility is to rely on application specific knowledge to guide compiler decisions. As it is impossible for a compiler alone to fully exploit such pieces of information. A possible approach to better utilize such knowledge is to put the programmers “in the loop”. Expert parallel programmers often have a good idea about coarse-grain parallelism and locality that they want to use for an application. On the other hand, fine-grain parallelism (e.g., ILP, SIMD) is tedious and specific to each underlying architecture, and is best left to the compiler. Furthermore, approximations will have opportunities to be refined using programmer knowledge. The key challenge is to create a programming environment where compiler techniques and programmer knowledge can be combined effectively. One of the difficulties is to design a usable interface between the compiler and the programmer.

3.3. Mathematical tools

All compilers have to deal with *sets* and relations. In classical compilers, these sets are finite: the set of statements of a program, the set of its variables, its abstract syntax tree (AST), its control-flow graph (CFG), and many others. It is only in the first phase of compilation, parsing, that one has to deal with infinite objects, regular and context-free languages, and those are represented by finite grammars, and are processed by a symbolic algorithm, `yacc` or one of its clones.

When tackling parallel programs and parallel compilation, it was soon realized that this position was no longer tenable. Since it makes no sense to ask whether a statement can be executed in parallel with itself, one has to consider sets of operations, which may be so large as to forbid an extensive representation, or even be infinite. The same is true for dependence sets, for memory cells, for communication sets, and for many other objects

a parallel compiler has to consider. The representation is to be *symbolic*, and all necessary algorithms have to be promoted to symbolic versions.

Such symbolic representations have to be efficient – the formula representing a set has to be much smaller than the set itself – and effective – the operations one needs, union, intersection, emptiness tests and many others – have to be feasible and fast. As a parenthesis, note that progress in algorithm design has blurred the distinction between polynomially-solvable and NP-complete problems, and between decidable and undecidable questions. For instance SAT, SMT, and ILP software tools solve efficiently many NP-complete problems, and the Z3 tool is able to "solve" many instances of the undecidable Hilbert's 10th problem.

Since the times of Pip and of the Polylib, Compsys has been active in the implementation of basic mathematical tools for program analysis and synthesis. Pip is still developed by Paul Feautrier and Cédric Bastoul, while the Polylib is now taken care of by the Inria Camus project, which introduced Ehrhart polynomials. These tools are still in use world-wide and they also have been reimplemented many times with (sometimes slight) improvements, e.g., as part of the Parma Polylib, of Sven Verdoolaege's Isl and Barvinok libraries, or of the Jollylib of Reservoir Labs. Other groups also made a lot of efforts towards the democratization of the use of polyhedral techniques, in particular the Alchemy Inria project, with Cloog and the development of Graphite in GCC, and Sadayappan's group in the USA, with the development of U. Bondhugula's Pluto prototype compiler. The same effort is made through the PPCG prototype compiler (for GPU) and Pencil (directives-based language on top of PPCG).

After 2009, Compsys continued to focus on the introduction of concepts and techniques to extend the polytope model, with a shift toward tools that may prepare the future. For instance, PoCo and C2fsm are able to parse general programs, not just SCoPs (static control programs), while the efficient handling of Boolean affine formulas [23] is a prerequisite for the construction of non-convex approximations. Euclidean lattices provide an efficient abstraction for the representation of spatial phenomena, and the construction of *critical lattices* as embedded in the tool Cl@k is a first step towards memory optimization in stream languages and may be useful in other situations. Our work on Chuba introduced a new element-wise array reuse analysis and the possibility of handling approximations. Our work on the analysis of while loops is both an extension of the polytope model itself (i.e., beyond SCoPs) and of its applications, here links with program termination and worst-case execution time (WCET) tools.

A recent example of the same approach is the proposal by Paul Feautrier to use polynomials for program analysis and optimization [5]. The associated tools are based on Handelman and Schweighofer theorems, the polynomial analogue of Farkas lemma. While this is definitely work in progress, with many unsolved questions, it has the potential of greatly enlarging the set of tractable programs.

As a last remark, observe that a common motif of these development is the transformation of finite algorithms into symbolic algorithms, able to solve very large or even infinite instances. For instance, PIP is a symbolic extension of the Simplex; our work on memory allocation is a symbolic extension of the familiar register allocation problem; loop scheduling extends DAG scheduling. Many other algorithms await their symbolic transformation: a case in point is resource-constrained scheduling.

CONVECS Project-Team

3. Research Program

3.1. New Formal Languages and their Concurrent Implementations

We aim at proposing and implementing new formal languages for the specification, implementation, and verification of concurrent systems. In order to provide a complete, coherent methodological framework, two research directions must be addressed:

- *Model-based specifications*: these are operational (i.e., constructive) descriptions of systems, usually expressed in terms of processes that execute concurrently, synchronize together and communicate. Process calculi are typical examples of model-based specification languages. The approach we promote is based on LOTOS NT (LNT for short), a formal specification language that incorporates most constructs stemming from classical programming languages, which eases its acceptance by students and industry engineers. LNT [35] is derived from the ISO standard E-LOTOS (2001), of which it represents the first successful implementation, based on a source-level translation from LNT to the former ISO standard LOTOS (1989). We are working both on the semantic foundations of LNT (enhancing the language with module interfaces and timed/probabilistic/stochastic features, compiling the m among n synchronization, etc.) and on the generation of efficient parallel and distributed code. Once equipped with these features, LNT will enable formally verified asynchronous concurrent designs to be implemented automatically.
- *Property-based specifications*: these are declarative (i.e., non-constructive) descriptions of systems, which express *what* a system should do rather than *how* the system should do it. Temporal logics and μ-calculi are typical examples of property-based specification languages. The natural models underlying value-passing specification languages, such as LNT, are Labeled Transition Systems (LTSs or simply *graphs*) in which the transitions between states are labeled by actions containing data values exchanged during handshake communications. In order to reason accurately about these LTSs, temporal logics involving data values are necessary. The approach we promote is based on MCL (*Model Checking Language*) [56], which extends the modal μ-calculus with data-handling primitives, fairness operators encoding generalized Büchi automata, and a functional-like language for describing complex transition sequences. We are working both on the semantic foundations of MCL (extending the language with new temporal and hybrid operators, translating these operators into lower-level formalisms, enhancing the type system, etc.) and also on improving the MCL on-the-fly model checking technology (devising new algorithms, enhancing ergonomy by detecting and reporting vacuity, etc.).

We address these two directions simultaneously, yet in a coherent manner, with a particular focus on applicable concurrent code generation and computer-aided verification.

3.2. Parallel and Distributed Verification

Exploiting large-scale high-performance computers is a promising way to augment the capabilities of formal verification. The underlying problems are far from trivial, making the correct design, implementation, fine-tuning, and benchmarking of parallel and distributed verification algorithms long-term and difficult activities. Sequential verification algorithms cannot be reused as such for this task: they are inherently complex, and their existing implementations reflect several years of optimizations and enhancements. To obtain good speedup and scalability, it is necessary to invent new parallel and distributed algorithms rather than to attempt a parallelization of existing sequential ones. We seek to achieve this objective by working along two directions:

- *Rigorous design:* Because of their high complexity, concurrent verification algorithms should themselves be subject to formal modeling and verification, as confirmed by recent trends in the certification of safety-critical applications. To facilitate the development of new parallel and distributed verification algorithms, we promote a rigorous approach based on formal methods and verification. Such algorithms will be first specified formally in LNT, then validated using existing model checking algorithms of the CADP toolbox. Second, parallel or distributed implementations of these algorithms will be generated automatically from the LNT specifications, enabling them to be experimented on large computing infrastructures, such as clusters and grids. As a side-effect, this "bootstrapping" approach would produce new verification tools that can later be used to self-verify their own design.
- *Performance optimization:* In devising parallel and distributed verification algorithms, particular care must be taken to optimize performance. These algorithms will face concurrency issues at several levels: grids of heterogeneous clusters (architecture-independence of data, dynamic load balancing), clusters of homogeneous machines connected by a network (message-passing communication, detection of stable states), and multi-core machines (shared-memory communication, thread synchronization). We will seek to exploit the results achieved in the parallel and distributed computing field to improve performance when using thousands of machines by reducing the number of connections and the messages exchanged between the cooperating processes carrying out the verification task. Another important issue is the generalization of existing LTS representations (explicit, implicit, distributed) in order to make them fully interoperable, such that compilers and verification tools can handle these models transparently.

3.3. Timed, Probabilistic, and Stochastic Extensions

Concurrent systems can be analyzed from a *qualitative* point of view, to check whether certain properties of interest (e.g., safety, liveness, fairness, etc.) are satisfied. This is the role of functional verification, which produces Boolean (yes/no) verdicts. However, it is often useful to analyze such systems from a *quantitative* point of view, to answer non-functional questions regarding performance over the long run, response time, throughput, latency, failure probability, etc. Such questions, which call for numerical (rather than binary) answers, are essential when studying the performance and dependability (e.g., availability, reliability, etc.) of complex systems.

Traditionally, qualitative and quantitative analyzes are performed separately, using different modeling languages and different software tools, often by distinct persons. Unifying these separate processes to form a seamless design flow with common modeling languages and analysis tools is therefore desirable, for both scientific and economic reasons. Technically, the existing modeling languages for concurrent systems need to be enriched with new features for describing quantitative aspects, such as probabilities, weights, and time. Such extensions have been well-studied and, for each of these directions, there exist various kinds of automata, e.g., discrete-time Markov chains for probabilities, weighted automata for weights, timed automata for hard real-time, continuous-time Markov chains for soft real-time with exponential distributions, etc. Nowadays, the next scientific challenge is to combine these individual extensions altogether to provide even more expressive models suitable for advanced applications.

Many such combinations have been proposed in the literature, and there is a large amount of models adding probabilities, weights, and/or time. However, an unfortunate consequence of this diversity is the confuse landscape of software tools supporting such models. Dozens of tools have been developed to implement theoretical ideas about probabilities, weights, and time in concurrent systems. Unfortunately, these tools do not interoperate smoothly, due both to incompatibilities in the underlying semantic models and to the lack of common exchange formats.

To address these issues, CONVECS follows two research directions:

- *Unifying the semantic models*. Firstly, we will perform a systematic survey of the existing semantic models in order to distinguish between their essential and non-essential characteristics, the goal being to propose a unified semantic model that is compatible with process calculi techniques for specifying and verifying concurrent systems. There are already proposals for unification either

theoretical (e.g., Markov automata) or practical (e.g., PRISM and MODEST modeling languages), but these languages focus on quantitative aspects and do not provide high-level control structures and data handling features (as LNT does, for instance). Work is therefore needed to unify process calculi and quantitative models, still retaining the benefits of both worlds.

- *Increasing the interoperability of analysis tools*. Secondly, we will seek to enhance the interoperability of existing tools for timed, probabilistic, and stochastic systems. Based on scientific exchanges with developers of advanced tools for quantitative analysis, we plan to evolve the CADP toolbox as follows: extending its perimeter of functional verification with quantitative aspects; enabling deeper connections with external analysis components for probabilistic, stochastic, and timed models; and introducing architectural principles for the design and integration of future tools, our long-term goal being the construction of a European collaborative platform encompassing both functional and non-functional analyzes.

3.4. Component-Based Architectures for On-the-Fly Verification

On-the-fly verification fights against state explosion by enabling an incremental, demand-driven exploration of LTSs, thus avoiding their entire construction prior to verification. In this approach, LTS models are handled implicitly by means of their *post* function, which computes the transitions going out of given states and thus serves as a basis for any forward exploration algorithm. On-the-fly verification tools are complex software artifacts, which must be designed as modularly as possible to enhance their robustness, reduce their development effort, and facilitate their evolution. To achieve such a modular framework, we undertake research in several directions:

- *New interfaces for on-the-fly LTS manipulation*. The current application programming interface (API) for on-the-fly graph manipulation, named OPEN/CAESAR [42], provides an "opaque" representation of states and actions (transitions labels): states are represented as memory areas of fixed size and actions are character strings. Although appropriate to the pure process algebraic setting, this representation must be generalized to provide additional information supporting an efficient construction of advanced verification features, such as: handling of the types, functions, data values, and parallel structure of the source program under verification, independence of transitions in the LTS, quantitative (timed/probabilistic/stochastic) information, etc.
- *Compositional framework for on-the-fly LTS analysis*. On-the-fly model checkers and equivalence checkers usually perform several operations on graph models (LTSs, Boolean graphs, etc.), such as exploration, parallel composition, partial order reduction, encoding of model checking and equivalence checking in terms of Boolean equation systems, resolution and diagnostic generation for Boolean equation systems, etc. To facilitate the design, implementation, and usage of these functionalities, it is necessary to encapsulate them in software components that could be freely combined and replaced. Such components would act as graph transformers, that would execute (on a sequential machine) in a way similar to coroutines and to the composition of lazy functions in functional programming languages. Besides its obvious benefits in modularity, such a component-based architecture will also make it possible to take advantage of multi-core processors.
- *New generic components for on-the-fly verification*. The quest for new on-the-fly components for LTS analysis must be pursued, with the goal of obtaining a rich catalog of interoperable components serving as building blocks for new analysis features. A long-term goal of this approach is to provide an increasingly large catalog of interoperable components covering all verification and analysis functionalities that appear to be useful in practice. It is worth noticing that some components can be very complex pieces of software (e.g., the encapsulation of an on-the-fly model checker for a rich temporal logic). Ideally, it should be possible to build a novel verification or analysis tool by assembling on-the-fly graph manipulation components taken from the catalog. This would provide a flexible means of building new verification and analysis tools by reusing generic, interoperable model manipulation components.

3.5. Real-Life Applications and Case Studies

We believe that theoretical studies and tool developments must be confronted with significant case studies to assess their applicability and to identify new research directions. Therefore, we seek to apply our languages, models, and tools for specifying and verifying formally real-life applications, often in the context of industrial collaborations.

CRYPT Team

3. Research Program

3.1. Public-Key Cryptanalysis

This project is interested in any public-key cryptanalysis, in the broad sense.

3.1.1. Mathematical Foundations

Historically, one useful side-effect of public-key cryptanalysis has been the introduction of advanced mathematical objects in cryptology, which were later used for cryptographic design. The most famous examples are elliptic curves (first introduced in cryptology to factor integer numbers), lattices (first introduced in cryptology to attack knapsack cryptosystems) and pairings over elliptic curves (first introduced in cryptology to attack the discrete logarithm problem over special elliptic curves). It is therefore interesting to develop the mathematics of public-key cryptanalysis. In particular, we would like to deepen our understanding of lattices by studying well-known mathematical aspects such as packing problems, transference theorems or random lattices.

3.1.2. Lattice Algorithms

Due to the strong interest surrounding lattice-based cryptography at the moment, our main focus is to attack lattice-based cryptosystems, particularly the most efficient ones (such as NTRU), and the ones providing new functionalities such as fully-homomorphic encryption or noisy multi-linear maps: recent cryptanalysis examples include [4], [5] for the latter, and [6] for the former. We want to assess the concrete security level of lattice-based cryptosystems, as has been done for cryptosystems based on integer factoring or discrete logarithms: this has been explored in [25], but needs to be developed. This requires to analyze and design the best algorithms for solving lattice problems, either exactly or approximately. In this area, much progress has been obtained the past few years (such as [26]), but we believe there is still more to come. We are working on new lattice computational records.

We are also interested in lattice-based cryptanalysis of non-lattice cryptosystems, by designing new attacks or improving old attacks. A well-known example is RSA for which the best attacks in certain settings are based on lattice techniques, following a seminal work by Coppersmith in 1996: recently [3], we improved the efficiency of some of these attacks on RSA, and we would like to extend this kind of results.

3.1.3. New Assumptions

In the past few years, new cryptographic functionalities (such as fully-homomorphic encryption, noisy multinear maps, indistinguishability obfuscation, etc.) have appeared, many of which being based on lattices. They usually introduce new algorithmic problems whose hardness is not well-understood. It is extremely important to study the hardness of these new assumptions, in order to evaluate the feasability of these new functionalities. Sometimes, the problem itself is not new, but the (aggressive) choices of parameters are: for instance, several implementations of fully-homorphic encryption used well-known lattice problems like LWE or BDD but with very large parameters which have not been studied much.

Currently, there are very few articles studying the concrete hardness of these new assumptions, especially compared to the articles using these new assumptions.

3.2. Secret-Key Cryptanalysis

Though secret-key cryptanalysis is the oldest form of cryptanalysis, there is regular progress in this area.

3.2.1. Hash Functions

In the past few years, the most important event has been the SHA-3 competition for a new hash function standard. This competition ended in 2012, with Keccak selected as the winner. We intend to study Keccak, together with the four other SHA-3 finalists. New cryptanalytical techniques designed to attack SHA-3 candidates are likely to be useful to attack other schemes. For instance, this was the case for the so-called rebound attack.

However, it is also interesting not to forget widespread hash functions: while it is now extremely easy to generate new MD5 collisions, a collision for SHA-1 has yet to be found, despite the existence of theoretical collision attacks faster than birthday attacks. Besides, there are still very few results on the SHA-2 standards family.

We may also be interested in related topics such as message authentication codes, especially those based on hash functions, which we explored in the past.

3.2.2. Symmetric Ciphers

Symmetric ciphers are widely deployed because of their high performances: a typical case is disk encryption and wireless communications.

We intend to study widespread block ciphers, such as the AES (now implemented in Intel processors) and Kasumi (used in UMTS) standards, as illustrated in recent publications [7], [9], [10], [9] of the team. Surprisingly, new attacks [24], [23] on the AES have appeared in the past few years, such as related-key attacks and single-key attacks. It is very important to find out if these attacks can be improved, even if they are very far from being practical. An interesting trend in block cipher cryptanalysis is to adapt recent attacks on hash functions: this is the reciprocal of the phenomenon of ten years ago, when Wang's MD5 collision attack was based on differential cryptanalysis.

Similarly to block ciphers, we intend to study widespread stream ciphers, such as RC4. The case of RC4 is particularly interesting due to the extreme simplicity of this cipher, and its deployment in numerous applications such as wireless Internet protocols. In the past few years, new attacks on RC4 based on various biases (such as [30]) have appeared, and several attacks on RC4 are used in WEP-attack tools.

DEDUCTEAM Exploratory Action

3. Research Program

3.1. From proof-checking to Interoperability

A new turn with Deduction modulo was taken when the idea of reasoning modulo an arbitrary equivalence relation was applied to typed λ-calculi with dependent types, that permits to express proofs as algorithms, using the Brouwer-Heyting-Kolmogorov interpretation and the Curry-de Bruijn-Howard correspondence [32]. It was shown in 2007, that extending the simplest λ-calculus with dependent types, the $\lambda\Pi$-calculus, with an equivalence relation, led to a calculus we called the $\lambda\Pi$-calculus modulo, that permitted to simulate many other λ-calculi, such as the Calculus of Constructions, designed to express proofs in specific theories.

This led to the development of a general proof-checker based on the $\lambda\Pi$-calculus modulo [3], that could be used to verify proofs coming from different proof systems, such as Coq [30], HOL [39], etc. To emphasize this versatility of our proof-system, we called it Dedukti —"to deduce" in Esperanto. This system is currently developed together with companion systems, Coqine, Holide, Focalide, and Zenonide, that permits to translate proofs from Coq, HOL, Focalize, and Zenon, to Dedukti. Other tools, such as Zenon Modulo, directly output proofs that can be checked by Dedukti.

Dedukti proofs can also be exported to other systems, in particular to the MMT format [47].

A thesis, which is at the root of our research effort, and which was already formulated by the team of the Logical Framework [38] is that proof-checkers should be theory independent. This is for instance expressed in the title of our invited talk at Icalp 2012: *A theory independent Curry-De Bruijn-Howard correspondence*.

Using a single prover to check proofs coming from different provers naturally led to investigate how these proofs could interact one with another. This issue is of prime importance because developments in proof systems are getting bigger and, unlike other communities in computer science, the proof-checking community has given little effort in the direction of standardization and interoperability. On a longer term we believe that, for each proof, we should be able to identify the systems in which it can be expressed.

3.2. Automated theorem proving

Deduction modulo has originally been proposed to solve a problem in automated theorem proving and some of the early work in this area focused on the design of an automated theorem proving method called *Resolution modulo*, but this method was so complex that it was never implemented. This method was simplified in 2010 [5] and it could then be implemented. This implementation that builds on the iProver effort [46] is called iProver modulo.

iProver modulo gave surprisingly good results [4], so that we use it now to search for proofs in many areas: in the theory of classes—also known as *B* set theory—, on finite structures, etc. Similar ideas have also been implemented for the tableau method with in particular several extensions of the *Zenon* automated theorem prover. More precisely, two extensions have been realized: the first one is called *Super Zenon* [13] [35] and is an extension to superdeduction (which is a variant of Deduction modulo), and the second one is called *Zenon Modulo* [33], [34] and is an extension to Deduction modulo. Both extensions have been extensively tested over first order problems (of the TPTP library), and also provide good results in terms of number of proved problems. In particular, these tools provide good performances in set theory, so that *Super Zenon* has been successfully applied to verify *B* proof rules of *Atelier B* (work in collaboration with *Siemens*). Similarly, we plan to apply *Zenon Modulo* in the framework of the *BWare* project to verify *B* proof obligations coming from the modeling of industrial applications.

More generally, we believe that proof-checking and automated theorem proving have a lot to learn from each other, because a proof is both a static linguistic object justifying the truth of a proposition and a dynamic process of proving this proposition.

3.3. Models of computation

The idea of Deduction modulo is that computation plays a major role in the foundations of mathematics. This led us to investigate the role played by computation in other sciences, in particular in physics. Some of this work can be seen as a continuation of Gandy's [36] on the fact that the physical Church-Turing thesis is a consequence of three principles of physics, two well-known: the homogeneity of space and time, and the existence of a bound on the velocity of information, and one more speculative: the existence of a bound on the density of information.

This led us to develop physically oriented models of computations.

DICE Team

3. Research Program

3.1. Introduction

Our aim is to address both

- challenges in the field of information technology, as well as
- trans-disciplinary issues emerging from the global impact of the digital revolution.

We believe that addressing both directions at the same time is an efficient way to be relevant in each of them.

We focus on intermediation platforms, which are becoming dominant systems in the Web industries. Intermediation platforms are systems which offer services to their users, which are well tuned for their expectation, thanks to the knowledge the platform has accumulated on usage. Search engines, social networks are examples of intermediation platforms. They ensure a gatekeeping function, always in direct contact to their users, providing them with the most relevant information or contact. Their economic model relies on a biface economy, with two types of users, one subsidizing the other. Their impact goes beyond the Web, and they disrupt step by step all sectors of the economy, transportation, Press, education, to name a few.

So far as IT is concerned, we focus on the technologies used for intermediation, which are at the basis of the largest online systems. For the transdisciplinary questions, we focus mostly on the new equilibrium that is resulting from the evolution of power balances due mostly to intermediation platforms.

3.2. Intermediation technologies

DICE focuses on intermediation platforms because of the central role they play in the new economy.

Intermediation platforms connect users to one another, or users to services with a very high accuracy. They rely on innovations both technological and social, which were unthinkable only ten years ago, when Facebook started. They allow communication and interaction between billions of users, gathered in the same digital space, both producers and consumers of data and services. State-of-the-art intermediation platforms include Facebook, Google, Twitter, GitHub, as well as Wikipedia, StackOverflow or Quora. These systems share a common design and their market penetration follows the same pattern. They are built around an initial minimal viable product based on a somehow naive low-tech implementation, which evolves after a few years of improvement to Web giants. Their domination now contributes to standardize the web industry, that means in particular:

- Gatekeeping, a direct relation with users together with services satisfying users' needs;
- Continuous data flows mapped to users' profiles;
- Search engines associating, in a relevant manner, producers, consumers and services.

These common characteristics lead to new software architectural standards, which are shared by all these systems, and used in the peripheral services developed in the ecosystem around their API:

- Authentication systems: openId, OAuth, ...
- Object graphs: opengraph, follower/followee scheme, ...
- DataFlow engines: Twitter storm, Google millwheel, ...
- Databases: noSql, keyValues stores, ...
- Web Browsers: javascript, dart, MEAN (Mongo, Express, Angular, Node),...

These architectural components impact the whole digital world. DICE targets systems that use standard architecture services but preserve some aspects we consider as disruptive ones: *data concentration, data symmetry* and *computational subsidiarity*. Our current research activity includes the following directions:

- Peer-to-peer design for preserving users' primary data;
- Third parties based organic systems providing subsidiary data computation hosted at peer sites;
- In-Browser applications that impact mobile device and demonstrate instantaneous usability;
- Flow-based computing enabling a stream based exchange of information between peers at runtime.

3.3. Economy of the digital world

The digital revolution is impacting all sectors of our societies and organizations, education, energy, transportation, health, to name a few. This revolution results in the phenomena of Schumpeter's *creative destruction*, with the disappearance of traditional sectors and the creation of new ones. Our societies, which did not anticipate the depth of the changes, have to struggle to adapt to the pace of the development of the industry. Legal reforms in various important sectors including taxation are at stake. Some countries, more reactive than others, are clearly pulling the changes, exploiting the benefits for businesses and the capacity to generate information and value, while others are trying to catch up with the global trends.

Data form the bricks of the information society, and their flows between users and services constitute the blood of the industry. We focus in DICE on the strategic role of data in this revolution, and in particular on the systems that harvest the data and concentrate it.

We are also interested in the global political impact of this revolution, which deeply changes the relations between governments and citizens. If the privacy is the focus of considerable attention, together with the state surveillance, in Europe in particular, it is only one aspect of the new knowledge made available. Social media produce considerable knowledge not only on individuals, but on populations as well, their economic fate, their political orientation, etc. On the other hand, open data from governments allow citizens to monitor the action of their governments, as well as to contribute to it. The digital revolution, with the capacity to access information in ways unthinkable in the recent past, modifies completely the balance of powers between citizens, states and corporations.

We investigate the digital world, and more precisely the power relations, from an interdisciplinary perspective. We simultaneously quantify power relations by studying data flows and the rise of intermediation platforms and produce an economical, political and ethical analysis of this new state of affairs. Namely, we show that areas such as the US or China dominate the digital world when others, such as Europe, do not succeed in proposing widely used intermediation platforms. This situation generates several conflicts between countries and companies and prevents weak countries from promoting their values and policies.

A new trend is emerging in the humanities, around in particular the digital studies, which promote the cooperation between computer scientists and specialists of social sciences. Among them, the Berkman center for Internet and Society in Harvard, the Medialab at MIT, or the Web Science Institute in the UK have gained strong visibility. They address positive as well as negative externalities of IT for societies, that is the new potentials offered as well as their risks. The Center for Information Technology Research in the Interest of Society in Berkeley also addresses fundamental political impacts on democracy, which can be enhanced by open data as well as another philosophy of political power as currently implemented in the State of California for instance. The Open Data Institute in the UK is also a leading center for political issues in Europe. France should catch up on these research trends, at the intersection of different scientific fields.

DREAMPAL Team

3. Research Program

3.1. New Models for New Technologies

Over the past 25 years there have been several hardware-architecture generations dedicated to massively parallel computing. We have contributed to them in the past, and shall continue doing so in the Dreampal project. The three generations, chronologically ordered, are:

- Supercomputers from the 80s and 90s, based on massively parallel architectures that are more or less distributed (from the Cray T3D or Connection Machine CM2 to GRID 5000). Computer scientists have proposed methods and tools for mapping sequential algorithms to those parallel architectures in order to extract maximum power from them. We have contributed in this area in the past: http://www.lifl.fr/west/team.html.
- Parallelism pervades the chips! A new challenge appears: hardware/software co-design, in order to obtain performance gains by designing algorithms together with the parallel architectures of chips adapted to the algorithms. During the previous decade many studies, including ours in the Inria DaRT team, were dedicated to this type of co-design. DaRT has contributed to the development of the OMG MARTE standard (http://www.omgmarte.org) and to its implementation on several parallel platforms. Gaspard2, our implementation of this concept, was identified as one of the key software tools developed at Inria: http://www.inria.fr/en/centre/lille/research/platforms-and-flagship-software/flagship-software.
- The new challenge of the 2010s is, in our opinion, the integration of dynamic reconfiguration and massive parallelism. New circuits with high-density integration and supporting dynamic hardware reconfiguration have been proposed. In such architectures one can dynamically change the architecture while an algorithm is running on it. The Dynamic Partial Reconfiguration (DPR) feature offered by recent FPGA boards even allows, in theory, to generate optimized hardware at runtime, by adding, removing, and replacing components on a by-need basis. This integration of dynamic reconfiguration and massive parallelism induces a new degree of complexity, which we, as computer scientists, need to understand and deal with order to make possible the design of applications running on such architectures. This is the main challenge that we address in the Dreampal project. We note that we adress these problems as computer scientists; we do, however, collaborate with electronics specialists in order to benefit from their expertise in 3-D FPGAs.

Excerpt from the HiPEAC vision 2011/12

> *"The advent of 3D stacking enables higher levels of integration and reduced costs for off-chip communications. The overall complexity is managed due to the separation in different dies, independently designed."*

FPGAs (Field Programmable Gate Arrays) are configurable circuits that have emerged as a privileged target platform for intensive signal processing applications. FPGAs take advantage of the latest technological developments in circuits. For example, the Virtex7 from Xilinx offers a 28-nanometer integration, which is only one or two generations behind the latest general-purpose processors. 3D-Stacked Integrated Circuits (3D SICs) consist of two or more conventional 2D circuits stacked on the top of each other and built into the same IC. Recently, 3D SICs have been released by Xilinx for the Virtex 7 FPGA family. 3D integration will vastly increase the integration capabilities of FPGA circuits. The convergence of massive parallelism and dynamic reconfiguration in inevitable: we believe it is one of the main challenges in computing for the current decade.

By incorporating the configuration and/or data/program memory on the top of the FPGA fabric, with fast and numerous connections between memory and elementary logic blocks (∼10000 connections between dies), it will be possible to obtain dynamically reconfigurable computing platforms with a very high reconfiguration rate. Such a rate was not possible before, due to the serial nature of the interface between the configuration memory and the FPGA fabric itself. The FPGA technology also enables massively parallel architectures due to the large number of programmable logic fabrics available on the chip. For instance, Xilinx demonstrated 3600 8-bit picoBlaze softcore processors running simultaneously on the Virtex-7 2000T FPGA. For specific applications, picoBlaze can be replaced by specialized hardware accelerators or other IPs (Intellectual Property) components. This opens the possibility of creating massively parallel IP-based machines.

3.2. Multi-softcore on 3D FPGA

From the 2010 Xilinx white paper on FPGAs:

> *"Unlike a processor, in which architecture of the ALU is fixed and designed in a general-purpose manner to execute various operations, the CLBs (configurable logic blocks) can be programmed with just the operations needed by the application... The FPGA architecture provides the flexibility to create a massive array of application-specific ALUs..The new solution enables high-bandwidth connectivity between multiple die by providing a much greater number of connections... enabling the integration of massive quantities of interconnect logic resources within a single package"*

Softcore processors are processors implemented using hardware synthesis. Proprietary solutions include PicoBlaze, MicroBlaze, Nios, and Nios II; open-source solutions include Leon, OpenRisk, and FC16. The choice is wide and many new solutions emerge, including multi-softcore implementations on FPGAs. An alternative to softcores are hardware accelerators on FPGAs, which are dedicated circuits that are an order of magnitude faster than softcores. Between these two approaches, there are other various approaches that connect IPs to softcores, in which, the processor's machine-code language is extended, and IP invocations become new instructions. We envisage a new class of softcores (we call them reflective softcores [0]), where almost everything is implemented in IPs; only the control flow is assigned to the softcore itself. The partial dynamic reconfiguration of next-generation FPGAs makes such dynamic IP management possible in practice. We believe that efficient reflective softcores on the new 3D-FPGAs should be as small as possible: low-performance generic hardware components (ALU, registers, memory, I/O...) should be replaced by dedicated high-performance IPs.

We are developing a sofcore processor called HoMade (http://www.lifl.fr/~dekeyser/Homade) following these ideas.

In the multi-reflective softcores that we develop, some softcores will be slaves and others will be masters. Massively parallel dynamically reconfigurable architectures of softcores can thus be envisaged. This requires, additionally, a parallel management of the partial dynamic reconfiguration system. This can be done, for example, on a given subset of softcores: a massively parallel reconfiguration will replace the current replication of a given IP with the replication of a new IP. Thanks to the new 3D-FPGAs this task can be performed efficiently and in parallel using the large number of 3D communication links (Through-Silicon-Vias). Our roadmap for HoMade is to evolve towards this multi-reflective softcore model.

3.3. When Hardware Meets Software

HIPEAC vision 2011/12: *"The number of cores and instruction set extensions increases with every new generation, requiring changes in the software to effectively exploit the new features."*

[0] Hereafter, by reflective system, we mean a system that is able to modify its own structure and behaviour while it is running. A reflective softcore thus dynamically adds, removes, and replaces IPs in the application running on it, and is able to dynamically modify its own program memory, thereby dynamically altering the program it is executing.

When the new massively parallel dynamically reconfigurable architectures become reality users will need languages for programming software applications on them. The languages will be themselves dynamic and parallel, in order to reflect and to fully exploit the dynamicity and parallelism of the architectures. Thus, developers will be able to invoke reconfiguration and call parallel instructions in their programs. This expressiveness comes with a cost, however, because new classes of bugs can be induced by the interaction between dynamic reconfiguration and parallelism; for example, deadlocks due to waiting for output from an IP that does not exist any more due to a reconfiguration. The detection and elimination of such bugs before deployment is paramount for cost-effectiveness and safety reasons.

Thus, we shall build an environment for developing software on parallel, dynamically reconfigurable architectures that will include languages and adequate formal analyses and verification tools for them, in addition to more traditional tools (emulators, compilers, etc). To this end we shall be using formal-semantics frameworks associated with easy-to-use formal verification tools in order to formally define our languages of interest and allow users to formally verify their programs. The K semantic framework (http://k-framework.org), developed jointly by Univs. Urbana Champaign, USA, and Iasi, Romania) is one such framework, which is mature enough (it has allowed defining a formal semantics of the largest subset of the C language to date, as well as many other languages from essentially all programming paradigms) and is familiar to us from previous work. In K, one can rapidly prototype a language definition and try several versions of the syntax and semantics of instructions. This is important in our project, where the proposed programming languages (in particular, the HoMade assembly language) will go through several versions before being stabilized. Moreover, once a language is defined in K one gets an interpreter of the language and one gains access to formal verification tools for free. We are also developing new analysis verification tools for K (in collaboration with the K team), which will be adapted and used in the Dreampal project.

ESTASYS Exploratory Action

3. Research Program

3.1. Systems of Systems, Heterogeneous Systems, Dynamicity, Statistical Model Checking

Formal methods rely on the notion of *transition system* (TS): an abstract machine that characterises a system's *complete* behaviour. This machine consists of a complete set of states (each representing full knowledge of the system at a given moment) and transitions between states, which may be labelled with labels chosen from some set of actions. This definition makes it necessary to have advanced knowledge of all the possible states of the system – to have a statically configured system. The algorithms used by formal methods perform an exhaustive exploration of the state space of the TS, so such methods suffer from the so-called *state-space explosion problem*. As a consequence, there are many real systems that are beyond the scope of such techniques. Despite this, over the last thirty years it has been shown that, when combined with heuristics such as partial order reductions or abstraction, **formal approaches are powerful enough to verify industrial-scale systems**.

The first wave of techniques was deployed to verify whether a certain set of (problem) states can be reached ('reachability'). Later, extensions of TS, such as *hybrid systems* and *stochastic automata*, were proposed to cope with new problems (e.g., energy consumption) or to reason on distributed real-time embedded components (possibly heterogeneous). It was quickly observed that the complexity of assessing correctness of such extended models arises not exclusively from the fact that they are large, but also because they introduce *undecidability*. As a concrete example, the reachability problem is already undecidable for any real-time system whose time evolution is described by a non-constant derivative equation.

This motivated the development of more efficient techniques that approximate the answer to the original problem or approximate the problem. Of these, perhaps the most successful quantitative technique is *Statistical Model Checking* , that can be seen as a trade-off between testing and formal verification. The core idea of SMC is to generate a number of *simulations* of the system and verify whether they satisfy a given property expressed in temporal logics, which can be done by using *runtime verification approaches*. The results are then used together with algorithms from the statistical area in order to decide whether the system satisfies the property with some probability. SMC resembles classical simulation-based techniques used in industry, but uses a formal model of systems and requirements. This not only gives a rigorous meaning to industrial practices, but also makes available more than twenty years of research in the area of *runtime verification*. Last but not least, **the use of statistical algorithms allows us to approximate undecidable problems**. Recent successful applications of SMC can be found in systems biology, security protocols and avionics. In particular, SMC was used to discover inconsistent requirements of an EADS airplane communication system.

3.1.1. Systems of Systems (SoS)

The advent of service-oriented and cloud architectures is leading to generations of computer systems that exhibit a new type of complexity: such systems are no longer statically configured, but comprise components that are systems in their own right, able to discover, select and bind on-the-fly to other components that can deliver services that they require. These complex systems, referred to as *Systems of Systems* (SoS), can change over time as each component creates and modifies the network over which it needs to operate: as they execute, the components create a network of their own and use it to fulfil their goals.

The Internet, made up of an unsupervised and rapidly growing, dynamically configured set of computers and physical connections, is an obvious illustration of the potential complexity of dynamic networks of interactions. Another example is the so-called "Flash Crash" in the U.S. equity market: on May 6, 2010, a block sale of 4.1 billion dollars of futures contracts executed on behalf of a fund-management company triggered a complex pattern of interactions between the high-frequency algorithmic trading systems (algos) that buy and sell blocks of financial instruments and made the Dow Jones Industrial Average drop more

than 600 points, representing the disappearance of 800 billion dollars of market value. This example is an illustration of the faulty divergence of SoS behaviour, where the system starts to misbehave and dynamically creates new components that follow the same pattern and make the problem worse. Examples of this include when a SoS detects high energy use and invokes a new component to reduce the energy, thus consuming *more* energy. **Until now, such divergence has been mostly handled by humans that eventually observe the faulty behaviour and manually intervene to stop it. This human-based solution is not always successful and clearly unsatisfactory, since it acts retrospectively, when the system has already failed.**

3.1.2. Grand Challenge and Breakthroughs of ESTASYS

SoS are an efficient means of achieving high performance and are thus becoming ubiquitous. Society's increasing reliance on SoS demands that they are reliable, but tools to guarantee this at the design stage do not exist. Most conventional formal analysis techniques, even those dedicated to adaptive systems, fail when applied to SoS because they are designed to reason on systems whose state space can be predicted in advance. **The grand challenge addressed by ESTASYS is the fundamental overhaul of formal methods techniques in the design of SoS life cycle**.

It is clear that SMC can be applied to the verification of complex systems. Unfortunately, SMC cannot yet be applied to SoS, because existing techniques are designed to capture the behaviour of statically configured systems, or systems whose dynamical configuration arises from permutations of known components. ESTASYS defines new abstract computational models and extend the state of the art of SMC to include SoS.

ESTASYS proposes a new formal methodology to support an evolutionary adaptive and iterative SoS life-cycle. *We foresee the following breakthroughs:*

1. Our ground-breaking computational model addresses the complex dynamic nature of SoS. The model is based on new interface theories that take into account behaviours of possibly unknown components and thus abstract what is unknown.
2. Cutting edge algorithms coming from the area of statistics and learning are exploited to make predictions about autonomous systems making local decisions. For example, **statistical abstraction** abstracts the behaviour of unknown environments; interleaving analysis and runtime monitoring of deployed systems to continuously update distributions embedded in the interfaces.
3. New statistical algorithms for SMC that scale efficiently and handle undecidability impacts the formal analysis of complex systems.
4. Our results are implemented in a professional toolset, ESTASYS-PLASMA, that is constructed in close collaboration with our industrial partners. This ensures relevance to industry and potentially high impact in the marketplace.

3.1.3. Methodology and Organization

ESTASYS's main challenge is to lay the foundation of a novel rigorous software construction methodology for SoS, based on simulation, statistics and industrial practices. ESTASYS establishes theories and empirical evidence for the introduction of formal verification-based approaches in the rigorous design of SoS.

ESTASYS addresses essential research questions for the introduction of formal techniques to support the SoS life-cycle. SoS occur in multiple disciplines and therefore there is a need for a common language. In particular, notions such as **autonomous decisions and dynamicity** must be standardized and well understood by those that will apply our methodology. Additionally, **characterizing the topological structure** of a SoS is essential for the study of component interactions and data exchanges. The complexity of SoS requires the development of a **sound formal semantic foundation** to support deployment of formal methods. We thus identify a minimal computational model that characterize SoS, on which classes of properties of interest can be defined. The project investigates new simulation-based approaches, combined with other domains (statistics, learning, ...), to verify such properties on the new computational model. Finally, ESTASYS identifies under which conditions the new techniques can be used, to take decisions during design and evolution time, leading to a fully integrated development cycle.

ESTASYS focuses on both the static and dynamic properties of SoS. ESTASYS establishes models for each component and investigates the connection and dynamical interactions between them. ESTASYS's activities are organize in six main tasks: tasks 1, 2 and 3 are responsible for breakthrough 1; task 4 is responsible for breakthrough 2; task 5 is responsible for breakthrough 3; task 6 is responsible for breakthrough 4.

Task 1. Characterizing SoS. Examples of SoS found in various areas, such as health care, smart buildings and energy grids, are analysed and used to standardize notions of autonomous decisions and dynamicity. We also study and classify SoS-related problems, such as faulty behaviour divergence. Our objective is to derive in Task 2 formal models that abstract the above classification.

Task 2. Formal Modeling of SoS. Classical theories do not provide for SoS, hence we require new formal models for SoS that take into account (i) dynamicity and emergent behaviours, (ii) autonomous decisions of components, and (iii) architectural constraints, including information regarding the viability of the hardware. In particular, we devise new logics tailored to the specific needs of SoS. Such logics, dynamic by nature, includes extended notions of quantification, such as energy, and considers hardware constraints and distributions of system configurations. Task 2 includes modelling the various components running within the SoS and their (dynamical) interactions. This requires the definition of a new type of interface able to work with heterogeneous components and to abstract the behaviour of unknown resources. Interfaces act as an abstraction for the internal behaviour of each component and encodes the dynamical constraints of the SoS. They are used to (i) model and define the authorised interactions between the components, (ii) reason on dynamical aspects and (iii) abstract unknown behaviour.

Task 3. Statistical abstraction interleaving design and deployment. Abstraction techniques are necessary to reduce the complexity of SoS and to model uncertainty. Specifically, **statistical abstractions** of the observed runtime behaviour of components is used to quantify, e.g., the probability that a number of new components satisfying some constraints is started at a given execution point. Runtime verification monitors the executions of the deployed system to create distributions embedded in the interfaces developed in Task 1. When a deployed system is available, ESTASYS interleaves simulation, analysis and runtime monitoring, using real behaviour to update the statistical abstractions, and eventually replace some of those abstractions by concrete ESTASYS-Interface models. The ESTASYS methodology adopts a Bayesian approach: (i) an initial, plausible distribution is 'guessed', based on whatever is known; (ii) the system is simulated using the current approximated distribution; (iii) the behaviour of the simulated system becomes the new approximation; (iv) the process is iterated as necessary. While learning-based simulation approaches, such as model fitting, can be used to learn the abstraction by conducting simulations from a finite set of initial components, we have to provide clear evidence that a global property holds on the system if it holds on its corresponding statistical abstraction. The task requires strong competences in statistics.

Task 4. Developing Efficient Simulation and Monitoring Algorithms for SoS. The ground-breaking models developed in Task 2 requires efficient simulation and monitoring techniques. This necessitates the study of new algorithms for dynamically configured systems and monitoring approaches to reason on heterogeneous components and the new quantitative logics and interface paradigms developed in Task 2.
A major difficulty in developing monitoring techniques for SoS is that the components have their own goals and behave differently in different environments. Unnecessary high-level hypotheses on properties may drastically increase simulation time and should be avoided.

Task 5. Developing Efficient Statistical Techniques for SoS. SoS pose new challenges for statistical techniques, requiring the study of new SMC algorithms dedicated to SoS goals. In contrast to existing SMC algorithms that can only be applied to pure stochastic systems, SMC algorithms for SoS have to take into account the non-deterministic aspects of autonomous decisions made by neighbour components. We postulate that this can be done by extending very recent advances in reinforcement learning algorithms. Rare events play an important role in system reliability, so we include rare-event simulation algorithms, such as importance sampling and importance splitting, which can reduce variance and significantly increase simulation efficiency.

Task 6. Evaluating the impact of statistical and simulation-based techniques. Evidence of the success of ESTASYS is provided by the publishing of a complete experimental environment, ESTASYS-PLASMA, that supports the empirical validation of ESTASYS's theories. ESTASYS-PLASMA contains efficient implementations of the results discovered in Tasks 2-5, and will provide intuitive feedback mechanisms so that the engineer can use the results of the verification process to improve SoS design.

GALAAD2 Team

3. Research Program

3.1. Introduction

Our scientific activity is structured according to three broad topics:

1. **Algebraic representations for geometric modeling**.
2. **Algebraic algorithms for geometric computing**,
3. **Symbolic-numeric methods for analysis**,

3.2. Algebraic representations for geometric modeling

Compact, efficient and structured descriptions of shapes are required in many scientific computations in engineering, such as "Isogeometric" Finite Elements methods, point cloud fitting problems or implicit surfaces defined by convolution. Our objective is to investigate new algebraic representations (or improve the existing ones) together with their analysis and implementations.

We are investigating representations, based on semi-algebraic models. Such non-linear models are able to capture efficiently complex shapes, using few data. However, they required specific methods to solve the underlying non-linear problems, which we are investigating.

Effective algebraic geometry is a natural framework for handling shape representations. This framework not only provides tools for modeling but it also allows to exploit rich geometric properties.

The above-mentioned tools of effective algebraic geometry make it possible to analyse in detail and separately algebraic varieties. We are interested in problems where collections of piecewise algebraic objects are involved. The properties of such geometrical structures are still not well controlled, and the traditional algorithmic geometry methods do not always extend to this context, which requires new investigations.

The use of piecewise algebraic representations also raises problems of approximation and reconstruction, on which we are working on. In this direction, we are studying B-spline function spaces with specified regularity associated to domain partitions.

Many geometric properties are, by nature, independent from the reference one chooses for performing analytic computations. This leads naturally to invariant theory. We are interested in exploiting these invariant properties, to develop compact and adapted representations of shapes.

3.3. Algebraic algorithms for geometric computing

This topic is directly related to polynomial system solving and effective algebraic geometry. It is our core expertise and many of our works are contributing to this area.

Our goal is to develop algebraic algorithms to efficiently perform geometric operations such as computing the intersection or self-intersection locus of algebraic surface patches, offsets, envelopes of surfaces, ...

The underlying representations behind the geometric models we consider are often of algebraic type. Computing with such models raises algebraic questions, which frequently appear as bottlenecks of the geometric problems.

In order to compute the solutions of a system of polynomial equations in several variables, we analyse and take advantage of the structure of the quotient ring defined by these polynomials. This raises questions of representing and computing normal forms in such quotient structures. The numerical and algebraic computations in this context lead us to study new approaches of normal form computations, generalizing the well-known Gröbner bases.

Geometric objects are often described in a parametric form. For performing efficiently on these objects, it can also be interesting to manipulate implicit representations. We consider particular projections techniques based on new resultant constructions or syzygies, which allow to transform parametric representations into implicit ones. These problems can be reformulated in terms of linear algebra. We investigate methods which exploit this matrix representation based on resultant constructions.

They involve structured matrices such as Hankel, Toeplitz, Bezoutian matrices or their generalization in several variables. We investigate algorithms that exploit their properties and their implications in solving polynomial equations.

We are also interested in the "effective" use of duality, that is, the properties of linear forms on the polynomials or quotient rings by ideals. We undertake a detailed study of these tools from an algorithmic perspective, which yields the answer to basic questions in algebraic geometry and brings a substantial improvement on the complexity of resolution of these problems.

We are also interested in subdivision methods, which are able to efficiently localise the real roots of polynomial equations. The specificities of these methods are local behavior, fast convergence properties and robustness. Key problems are related to the analysis of multiple points.

An important issue while developing these methods is to analyse their practical and algorithmic behavior. Our aim is to obtain good complexity bounds and practical efficiency by exploiting the structure of the problem.

3.4. Symbolic numeric analysis

While treating practical problems, noisy data appear and incertitude has to be taken into account. The objective is to devise adapted techniques for analyzing the geometric properties of the algebraic models in this context.

Analysing a geometric model requires tools for structuring it, which first leads to study its singularities and its topology. In many contexts, the input representation is given with some error so that the analysis should take into account not only one model but a neighborhood of models.

The analysis of singularities of geometric models provides a better understanding of their structures. As a result, it may help us better apprehend and approach modeling problems. We are particularly interested in applying singularity theory to cases of implicit curves and surfaces, silhouettes, shadows curves, moved curves, medial axis, self-intersections, appearing in algorithmic problems in CAGD and shape analysis.

The representation of such shapes is often given with some approximation error. It is not surprising to see that symbolic and numeric computations are closely intertwined in this context. Our aim is to exploit the complementarity of these domains, in order to develop controlled methods.

The numerical problems are often approached locally. However, in many situations it is important to give global answers, making it possible to certify computation. The symbolic-numeric approach combining the algebraic and analytical aspects, intends to address these local-global problems. Especially, we focus on certification of geometric predicates that are essential for the analysis of geometrical structures.

The sequence of geometric constructions, if treated in an exact way, often leads to a rapid complexification of the problems. It is then significant to be able to approximate the geometric objects while controlling the quality of approximation. We investigate subdivision techniques based on the algebraic formulation of our problems which allow us to control the approximation, while locating interesting features such as singularities.

According to an engineer in CAGD, the problems of singularities obey the following rule: less than 20% of the treated cases are singular, but more than 80% of time is necessary to develop a code allowing to treat them correctly. Degenerated cases are thus critical from both theoretical and practical perspectives. To resolve these difficulties, in addition to the qualitative studies and classifications, we also study methods of *perturbations* of symbolic systems, or adaptive methods based on exact arithmetics.

The problem of decomposition and factorisation is also important. We are interested in a new type of algorithms that combine the numerical and symbolic aspects, and are simultaneously more effective and reliable. A typical problem in this direction is the problem of approximate factorization, which requires to analyze perturbations of the data, which enables us to break up the problem.

GALLIUM Project-Team

3. Research Program

3.1. Programming languages: design, formalization, implementation

Like all languages, programming languages are the media by which thoughts (software designs) are communicated (development), acted upon (program execution), and reasoned upon (validation). The choice of adequate programming languages has a tremendous impact on software quality. By "adequate", we mean in particular the following four aspects of programming languages:

- **Safety.** The programming language must not expose error-prone low-level operations (explicit memory deallocation, unchecked array accesses, etc) to the programmers. Further, it should provide constructs for describing data structures, inserting assertions, and expressing invariants within programs. The consistency of these declarations and assertions should be verified through compile-time verification (e.g. static type checking) and run-time checks.
- **Expressiveness.** A programming language should manipulate as directly as possible the concepts and entities of the application domain. In particular, complex, manual encodings of domain notions into programmatic notations should be avoided as much as possible. A typical example of a language feature that increases expressiveness is pattern matching for examination of structured data (as in symbolic programming) and of semi-structured data (as in XML processing). Carried to the extreme, the search for expressiveness leads to domain-specific languages, customized for a specific application area.
- **Modularity and compositionality.** The complexity of large software systems makes it impossible to design and develop them as one, monolithic program. Software decomposition (into semi-independent components) and software composition (of existing or independently-developed components) are therefore crucial. Again, this modular approach can be applied to any programming language, given sufficient fortitude by the programmers, but is much facilitated by adequate linguistic support. In particular, reflecting notions of modularity and software components in the programming language enables compile-time checking of correctness conditions such as type correctness at component boundaries.
- **Formal semantics.** A programming language should fully and formally specify the behaviours of programs using mathematical semantics, as opposed to informal, natural-language specifications. Such a formal semantics is required in order to apply formal methods (program proof, model checking) to programs.

Our research work in language design and implementation centers around the statically-typed functional programming paradigm, which scores high on safety, expressiveness and formal semantics, complemented with full imperative features and objects for additional expressiveness, and modules and classes for compositionality. The OCaml language and system embodies many of our earlier results in this area [48]. Through collaborations, we also gained experience with several domain-specific languages based on a functional core, including distributed programming (JoCaml), XML processing (XDuce, CDuce), reactive functional programming, and hardware modeling.

3.2. Type systems

Type systems [65] are a very effective way to improve programming language reliability. By grouping the data manipulated by the program into classes called types, and ensuring that operations are never applied to types over which they are not defined (e.g. accessing an integer as if it were an array, or calling a string as if it were a function), a tremendous number of programming errors can be detected and avoided, ranging from the trivial (misspelled identifier) to the fairly subtle (violation of data structure invariants). These restrictions are also very effective at thwarting basic attacks on security vulnerabilities such as buffer overflows.

The enforcement of such typing restrictions is called type checking, and can be performed either dynamically (through run-time type tests) or statically (at compile-time, through static program analysis). We favor static type checking, as it catches bugs earlier and even in rarely-executed parts of the program, but note that not all type constraints can be checked statically if static type checking is to remain decidable (i.e. not degenerate into full program proof). Therefore, all typed languages combine static and dynamic type-checking in various proportions.

Static type checking amounts to an automatic proof of partial correctness of the programs that pass the compiler. The two key words here are *partial*, since only type safety guarantees are established, not full correctness; and *automatic*, since the proof is performed entirely by machine, without manual assistance from the programmer (beyond a few, easy type declarations in the source). Static type checking can therefore be viewed as the poor man's formal methods: the guarantees it gives are much weaker than full formal verification, but it is much more acceptable to the general population of programmers.

3.2.1. Type systems and language design.

Unlike most other uses of static program analysis, static type-checking rejects programs that it cannot analyze safe. Consequently, the type system is an integral part of the language design, as it determines which programs are acceptable and which are not. Modern typed languages go one step further: most of the language design is determined by the *type structure* (type algebra and typing rules) of the language and intended application area. This is apparent, for instance, in the XDuce and CDuce domain-specific languages for XML transformations [59], [53], whose design is driven by the idea of regular expression types that enforce DTDs at compile-time. For this reason, research on type systems – their design, their proof of semantic correctness (type safety), the development and proof of associated type checking and inference algorithms – plays a large and central role in the field of programming language research, as evidenced by the huge number of type systems papers in conferences such as Principles of Programming Languages.

3.2.2. Polymorphism in type systems.

There exists a fundamental tension in the field of type systems that drives much of the research in this area. On the one hand, the desire to catch as many programming errors as possible leads to type systems that reject more programs, by enforcing fine distinctions between related data structures (say, sorted arrays and general arrays). The downside is that code reuse becomes harder: conceptually identical operations must be implemented several times (say, copying a general array and a sorted array). On the other hand, the desire to support code reuse and to increase expressiveness leads to type systems that accept more programs, by assigning a common type to broadly similar objects (for instance, the `Object` type of all class instances in Java). The downside is a loss of precision in static typing, requiring more dynamic type checks (downcasts in Java) and catching fewer bugs at compile-time.

Polymorphic type systems offer a way out of this dilemma by combining precise, descriptive types (to catch more errors statically) with the ability to abstract over their differences in pieces of reusable, generic code that is concerned only with their commonalities. The paradigmatic example is parametric polymorphism, which is at the heart of all typed functional programming languages. Many forms of polymorphic typing have been studied since then. Taking examples from our group, the work of Rémy, Vouillon and Garrigue on row polymorphism [69], integrated in OCaml, extended the benefits of this approach (reusable code with no loss of typing precision) to object-oriented programming, extensible records and extensible variants. Another example is the work by Pottier on subtype polymorphism, using a constraint-based formulation of the type system [66]. Finally, the notion of "coercion polymorphism" proposed by Cretin and Rémy [28] combines and generalizes both parametric and subtyping polymorphism.

3.2.3. Type inference.

Another crucial issue in type systems research is the issue of type inference: how many type annotations must be provided by the programmer, and how many can be inferred (reconstructed) automatically by the typechecker? Too many annotations make the language more verbose and bother the programmer with unnecessary details. Too few annotations make type checking undecidable, possibly requiring heuristics,

which is unsatisfactory. OCaml requires explicit type information at data type declarations and at component interfaces, but infers all other types.

In order to be predictable, a type inference algorithm must be complete. That is, it must not find *one*, but *all* ways of filling in the missing type annotations to form an explicitly typed program. This task is made easier when all possible solutions to a type inference problem are *instances* of a single, *principal* solution.

Maybe surprisingly, the strong requirements – such as the existence of principal types – that are imposed on type systems by the desire to perform type inference sometimes lead to better designs. An illustration of this is row variables. The development of row variables was prompted by type inference for operations on records. Indeed, previous approaches were based on subtyping and did not easily support type inference. Row variables have proved simpler than structural subtyping and more adequate for typechecking record update, record extension, and objects.

Type inference encourages abstraction and code reuse. A programmer's understanding of his own program is often initially limited to a particular context, where types are more specific than strictly required. Type inference can reveal the additional generality, which allows making the code more abstract and thus more reuseable.

3.3. Compilation

Compilation is the automatic translation of high-level programming languages, understandable by humans, to lower-level languages, often executable directly by hardware. It is an essential step in the efficient execution, and therefore in the adoption, of high-level languages. Compilation is at the interface between programming languages and computer architecture, and because of this position has had considerable influence on the designs of both. Compilers have also attracted considerable research interest as the oldest instance of symbolic processing on computers.

Compilation has been the topic of much research work in the last 40 years, focusing mostly on high-performance execution ("optimization") of low-level languages such as Fortran and C. Two major results came out of these efforts: one is a superb body of performance optimization algorithms, techniques and methodologies; the other is the whole field of static program analysis, which now serves not only to increase performance but also to increase reliability, through automatic detection of bugs and establishment of safety properties. The work on compilation carried out in the Gallium group focuses on a less investigated topic: compiler certification.

3.3.1. Formal verification of compiler correctness.

While the algorithmic aspects of compilation (termination and complexity) have been well studied, its semantic correctness – the fact that the compiler preserves the meaning of programs – is generally taken for granted. In other terms, the correctness of compilers is generally established only through testing. This is adequate for compiling low-assurance software, themselves validated only by testing: what is tested is the executable code produced by the compiler, therefore compiler bugs are detected along with application bugs. This is not adequate for high-assurance, critical software which must be validated using formal methods: what is formally verified is the source code of the application; bugs in the compiler used to turn the source into the final executable can invalidate the guarantees so painfully obtained by formal verification of the source.

To establish strong guarantees that the compiler can be trusted not to change the behavior of the program, it is necessary to apply formal methods to the compiler itself. Several approaches in this direction have been investigated, including translation validation, proof-carrying code, and type-preserving compilation. The approach that we currently investigate, called *compiler verification*, applies program proof techniques to the compiler itself, seen as a program in particular, and use a theorem prover (the Coq system) to prove that the generated code is observationally equivalent to the source code. Besides its potential impact on the critical software industry, this line of work is also scientifically fertile: it improves our semantic understanding of compiler intermediate languages, static analyses and code transformations.

3.4. Interface with formal methods

Formal methods refer collectively to the mathematical specification of software or hardware systems and to the verification of these systems against these specifications using computer assistance: model checkers, theorem provers, program analyzers, etc. Despite their costs, formal methods are gaining acceptance in the critical software industry, as they are the only way to reach the required levels of software assurance.

In contrast with several other Inria projects, our research objectives are not fully centered around formal methods. However, our research intersects formal methods in the following two areas, mostly related to program proofs using proof assistants and theorem provers.

3.4.1. Software-proof codesign

The current industrial practice is to write programs first, then formally verify them later, often at huge costs. In contrast, we advocate a codesign approach where the program and its proof of correctness are developed in interaction, and are interested in developing ways and means to facilitate this approach. One possibility that we currently investigate is to extend functional programming languages such as Caml with the ability to state logical invariants over data structures and pre- and post-conditions over functions, and interface with automatic or interactive provers to verify that these specifications are satisfied. Another approach that we practice is to start with a proof assistant such as Coq and improve its capabilities for programming directly within Coq.

3.4.2. Mechanized specifications and proofs for programming languages components

We emphasize mathematical specifications and proofs of correctness for key language components such as semantics, type systems, type inference algorithms, compilers and static analyzers. These components are getting so large that machine assistance becomes necessary to conduct these mathematical investigations. We have already mentioned using proof assistants to verify compiler correctness. We are also interested in using them to specify and reason about semantics and type systems. These efforts are part of a more general research topic that is gaining importance: the formal verification of the tools that participate in the construction and certification of high-assurance software.

GCG Team

3. Research Program

3.1. Foundations

It has been ten years now since Intel bumped on the energy wall. Parallelism is now ubiquitous, not only restricted to expensive servers dedicated to some regular scientific computation. Also, the panel of possible mainstream architectures became extremely diverse. The use of byte-codes (e.g. nVIDIA PTX) along with Just-In-Time (JIT) compilation allowed fast evolution of designs. Quite recently, silicon companies understood that this heterogeneity should be integrated into the same chip (e.g. ARM big.LITTLE, nVIDIA Tegra K1); also re-configurable architectures (from FPGA to CGRA) are becoming present in such design as specialization is clearly useful to increase performance with less increase in energy consumption. Even cache-size, crossbar will be dynamically re-configurable; distributed DVFS being now mainstream... Postponing the decision of where and how (depending on the context) to execute part of an application, involves the use of late/adaptive compilation so as to avoid code size blowing. This observation is amplified by the fact that application behavior gets more and more dominated by data-characteristics. This is precisely what motivated more than fifteen years ago the development of dynamic compiler optimization technology. Many transformations, decisions, code-generation phases done by a compiler are now critically required to be postponed at run-time when the information is becoming available. But, this is not to mention the need of auto-tuning and adaptive compilation that imposes itself to address the increasing complexity (and hard to model) of each individual core.

The research direction of GCG is motivated by the perspective of optimizing (sometimes complex and irregular) micro-kernels for a single core (SIMD/VLIW). It starts from the observation that despite the clear motivation for JIT/dynamic compilation, despite its clear maturity, we lost the battle of performance portability: such technologies are not as optimizing as we pretended it would be. The reason for this defeat is that there is no perfect place to analyze, optimize, transform. On one hand “ JIT-ing” source-level code would usually be too slow, while on the other hand byte-code close to machine-level lost high-level semantics. Apart from spending its time to retrieve somehow obvious information, the JIT-compiler has to deal with limited resources, with realistic time constraints. Thus the need for being hybrid, in other words combine static and dynamic compilation/analysis techniques using rich intermediate languages.

Hybrid compilation consists in combining in any possible ways static analysis with profiling and run-time tests, but also ahead-of-time with run-time code optimization. This leads GCG to put efforts on researches on hybrid compilation frameworks but also on compiler architecture design. This last is to address the difficult problem of information telescoping (maintain of information of different type) and the problem of code size.

Current projects include:

- characterization of applications (I/O complexity) and profiling feedback using trace analyses;
- combined scheduling and memory allocation for irregular applications;
- extension of the polyhedral model using hybrid analysis and compilation;
- design, promotion and development of an hybrid and extensible byte-code, Tirex;
- design of a run-time handling communications, scheduling and placement for distributed memory parallel architectures.

GEOMETRICA Project-Team

3. Research Program

3.1. Mesh Generation and Geometry Processing

Meshes are becoming commonplace in a number of applications ranging from engineering to multimedia through biomedecine and geology. For rendering, the quality of a mesh refers to its approximation properties. For numerical simulation, a mesh is not only required to faithfully approximate the domain of simulation, but also to satisfy size as well as shape constraints. The elaboration of algorithms for automatic mesh generation is a notoriously difficult task as it involves numerous geometric components: Complex data structures and algorithms, surface approximation, robustness as well as scalability issues. The recent trend to reconstruct domain boundaries from measurements adds even further hurdles. Armed with our experience on triangulations and algorithms, and with components from the CGAL library, we aim at devising robust algorithms for 2D, surface, 3D mesh generation as well as anisotropic meshes. Our research in mesh generation primarily focuses on the generation of simplicial meshes, i.e. triangular and tetrahedral meshes. We investigate both greedy approaches based upon Delaunay refinement and filtering, and variational approaches based upon energy functionals and associated minimizers.

The search for new methods and tools to process digital geometry is motivated by the fact that previous attempts to adapt common signal processing methods have led to limited success: Shapes are not just another signal but a new challenge to face due to distinctive properties of complex shapes such as topology, metric, lack of global parameterization, non-uniform sampling and irregular discretization. Our research in geometry processing ranges from surface reconstruction to surface remeshing through curvature estimation, principal component analysis, surface approximationand surface mesh parameterization. Another focus is on the robustness of the algorithms to defect-laden data. This focus stems from the fact that acquired geometric data obtained through measurements or designs are rarely usable directly by downstream applications. This generates bottlenecks, i.e., parts of the processing pipeline which are too labor-intensive or too brittle for practitioners. Beyond reliability and theoretical foundations, our goal is to design methods which are also robust to raw, unprocessed inputs.

3.2. Topological and Geometric Inference

Due to the fast evolution of data acquisition devices and computational power, scientists in many areas are asking for efficient algorithmic tools for analyzing, manipulating and visualizing more and more complex shapes or complex systems from approximative data. Many of the existing algorithmic solutions which come with little theoretical guarantee provide unsatisfactory and/or unpredictable results. Since these algorithms take as input discrete geometric data, it is mandatory to develop concepts that are rich enough to robustly and correctly approximate continuous shapes and their geometric properties by discrete models. Ensuring the correctness of geometric estimations and approximations on discrete data is a sensitive problem in many applications.

Data sets being often represented as point sets in high dimensional spaces, there is a considerable interest in analyzing and processing data in such spaces. Although these point sets usually live in high dimensional spaces, one often expects them to be located around unknown, possibly non linear, low dimensional shapes. These shapes are usually assumed to be smooth submanifolds or more generally compact subsets of the ambient space. It is then desirable to infer topological (dimension, Betti numbers,...) and geometric characteristics (singularities, volume, curvature,...) of these shapes from the data. The hope is that this information will help to better understand the underlying complex systems from which the data are generated. In spite of recent promising results, many problems still remain open and to be addressed, need a tight collaboration between mathematicians and computer scientists. In this context, our goal is to contribute to the development of new mathematically well founded and algorithmically efficient geometric tools for data analysis and processing of complex geometric objects. Our main targeted areas of application include machine learning, data mining, statistical analysis, and sensor networks.

3.3. Data Structures and Robust Geometric Computation

GEOMETRICA has a large expertise of algorithms and data structures for geometric problems.We are pursuing efforts to design efficient algorithms from a theoretical point of view, but we also put efforts in the effective implementation of these results.

In the past years, we made significant contributions to algorithms for computing Delaunay triangulations (which are used by meshes in the above paragraph). We are still working on the practical efficiency of existing algorithms to compute or to exploit classical Euclidean triangulations in 2 and 3 dimensions, but the current focus of our research is more aimed towards extending the triangulation efforts in several new directions of research.

One of these directions is the triangulation of non Euclidean spaces such as periodic or projective spaces, with various potential applications ranging from astronomy to granular material simulation.

Another direction is the triangulation of moving points, with potential applications to fluid dynamics where the points represent some particles of some evolving physical material, and to variational methods devised to optimize point placement for meshing a domain with a high quality elements.

Increasing the dimension of space is also a stimulating direction of research, as triangulating points in medium dimension (say 4 to 15) has potential applications and raises new challenges to trade exponential complexity of the problem in the dimension for the possibility to reach effective and practical results in reasonably small dimensions.

On the complexity analysis side, we pursue efforts to obtain complexity analysis in some practical situations involving randomized or stochastic hypotheses. On the algorithm design side, we are looking for new paradigms to exploit parallelism on modern multicore hardware architectures.

Finally, all this work is done while keeping in mind concerns related to effective implementation of our work, practical efficiency and robustness issues which have become a background task of all different works made by GEOMETRICA.

GRACE Project-Team

3. Research Program

3.1. Algorithmic Number Theory

Algorithmic Number Theory is concerned with replacing special cases with general algorithms to solve problems in number theory. In the Grace project, it appears in three main threads:

- fundamental algorithms for integers and polynomials (including primality and factorization);
- algorithms for finite fields (including discrete logarithms); and
- algorithms for algebraic curves.

Clearly, we use computer algebra in many ways. Research in cryptology has motivated a renewed interest in Algorithmic Number Theory in recent decades—but the fundamental problems still exist *per se*. Indeed, while algorithmic number theory application in cryptanalysis is epitomized by applying factorization to breaking RSA public key, many other problems, are relevant to various area of computer science. Roughly speaking, the problems of the cryptological world are of bounded size, whereas Algorithmic Number Theory is also concerned with asymptotic results.

3.2. Arithmetic Geometry: Curves and their Jacobians

Arithmetic Geometry is the meeting point of algebraic geometry and number theory: that is, the study of geometric objects defined over arithmetic number systems (such as the integers and finite fields). The fundamental objects for our applications in both coding theory and cryptology are curves and their Jacobians over finite fields.

An algebraic *plane curve*$\mathcal{X}$ over a field $\mathbf{K}$ is defined by an equation

$$\mathcal{X} : F_{\mathcal{X}}(x, y) = 0 \quad \text{where } F_{\mathcal{X}} \in \mathbf{K}[x, y].$$

(Not every curve is planar—we may have more variables, and more defining equations—but from an algorithmic point of view, we can always reduce to the plane setting.) The *genus* $g_{\mathcal{X}}$ of $\mathcal{X}$ is a non-negative integer classifying the essential geometric complexity of $\mathcal{X}$; it depends on the degree of $F_{\mathcal{X}}$ and on the number of singularities of $\mathcal{X}$. The simplest curves with nontrivial Jacobians are curves of genus 1, known as *elliptic curves*; they are typically defined by equations of the form $y^2 = x^3 + Ax + B$. Elliptic curves are particularly important given their central role in public-key cryptography over the past two decades. Curves of higher genus are important in both cryptography and coding theory.

The curve $\mathcal{X}$ is associated in a functorial way with an algebraic group $J_{\mathcal{X}}$, called the *Jacobian* of $\mathcal{X}$. The group $J_{\mathcal{X}}$ has a geometric structure: its elements correspond to points on a $g_{\mathcal{X}}$-dimensional projective algebraic group variety. Typically, we do not compute with the equations defining this projective variety: there are too many of them, in too many variables, for this to be convenient. Instead, we use fast algorithms based on the representation in terms of classes of formal sums of points on $\mathcal{X}$.

3.3. Curve-Based cryptology

Jacobians of curves are excellent candidates for cryptographic groups when constructing efficient instances of public-key cryptosystems. Diffie–Hellman key exchange is an instructive example.

Suppose Alice and Bob want to establish a secure communication channel. Essentially, this means establishing a common secret *key*, which they will then use for encryption and decryption. Some decades ago, they would have exchanged this key in person, or through some trusted intermediary; in the modern, networked world, this is typically impossible, and in any case completely unscalable. Alice and Bob may be anonymous parties who want to do e-business, for example, in which case they cannot securely meet, and they have no way to be sure of each other's identities. Diffie–Hellman key exchange solves this problem. First, Alice and Bob publicly agree on a cryptographic group G with a generator P (of order N); then Alice secretly chooses an integer a from $[1..N]$, and sends aP to Bob. In the meantime, Bob secretly chooses an integer b from $[1..N]$, and sends bP to Alice. Alice then computes $a(bP)$, while Bob computes $b(aP)$; both have now computed abP, which becomes their shared secret key. The security of this key depends on the difficulty of computing abP given P, aP, and bP; this is the Computational Diffie–Hellman Problem (CDHP). In practice, the CDHP corresponds to the Discrete Logarithm Problem (DLP), which is to determine a given P and aP.

This simple protocol has been in use, with only minor modifications, since the 1970s. The challenge is to create examples of groups G with a relatively compact representation and an efficiently computable group law, and such that the DLP in G is hard (ideally approaching the exponential difficulty of the DLP in an abstract group). The Pohlig–Hellman reduction shows that the DLP in G is essentially only as hard as the DLP in its largest prime-order subgroup. We therefore look for compact and efficient groups of prime order.

The classic example of a group suitable for the Diffie–Hellman protocol is the multiplicative group of a finite field $\mathbf{F}_q$. There are two problems that render its usage somewhat less than ideal. First, it has too much structure: we have a subexponential Index Calculus attack on the DLP in this group, so while it is very hard, the DLP falls a long way short of the exponential difficulty of the DLP in an abstract group. Second, there is only one such group for each q: its subgroup treillis depends only on the factorization of $q-1$, and requiring $q-1$ to have a large prime factor eliminates many convenient choices of q.

This is where Jacobians of algebraic curves come into their own. First, elliptic curves and Jacobians of genus 2 curves do not have a subexponential index calculus algorithm: in particular, from the point of view of the DLP, a generic elliptic curve is currently *as strong as* a generic group of the same size. Second, they provide some diversity: we have many degrees of freedom in choosing curves over a fixed $\mathbf{F}_q$, with a consequent diversity of possible cryptographic group orders. Furthermore, an attack which leaves one curve vulnerable may not necessarily apply to other curves. Third, viewing a Jacobian as a geometric object rather than a pure group allows us to take advantage of a number of special features of Jacobians. These features include efficiently computable pairings, geometric transformations for optimised group laws, and the availability of efficiently computable non-integer endomorphisms for accelerated encryption and decryption.

3.4. Algebraic Coding Theory

Coding Theory studies originated with the idea of using redundancy in messages to protect against noise and errors. The last decade of the 20th century has seen the success of so-called iterative decoding methods, which enable us to get very close to the Shannon capacity. The capacity of a given channel is the best achievable transmission *rate* for reliable transmission. The consensus in the community is that this capacity is more easily reached with these iterative and probabilistic methods than with algebraic codes (such as Reed–Solomon codes).

However, algebraic coding is useful in settings other than the Shannon context. Indeed, the Shannon setting is a random case setting, and promises only a vanishing error probability. In contrast, the algebraic Hamming approach is a worst case approach: under combinatorial restrictions on the noise, the noise can be adversarial, with strictly zero errors.

These considerations are renewed by the topic of *list decoding* after the breakthrough of Guruswami and Sudan at the end of the nineties. List decoding relaxes the uniqueness requirement of decoding, allowing a small list of candidates to be returned instead of a single codeword. List decoding can reach a capacity close to the Shannon capacity, with zero failure, with small lists, in the adversarial case. The method of Guruswami and Sudan enabled list decoding of most of the main algebraic codes: Reed–Solomon codes and

Algebraic–Geometry (AG) codes and new related constructions "capacity-achieving list decodable codes". These results open the way to applications again adversarial channels, which correspond to worst case settings in the classical computer science language.

Another avenue of our studies is AG codes over various geometric objects. Although Reed–Solomon codes are the best possible codes for a given alphabet, they are very limited in their length, which cannot exceed the size of the alphabet. AG codes circumvent this limitation, using the theory of algebraic curves over finite fields to construct long codes over a fixed alphabet. The striking result of Tsfasman–Vladut–Zink showed that codes better than random codes can be built this way, for medium to large alphabets. Disregarding the asymptotic aspects and considering only finite length, AG codes can be used either for longer codes with the same alphabet, or for codes with the same length with a smaller alphabet (and thus faster underlying arithmetic).

From a broader point of view, wherever Reed–Solomon codes are used, we can substitute AG codes with some benefits: either beating random constructions, or beating Reed–Solomon codes which are of bounded length for a given alphabet.

Another area of Algebraic Coding Theory with which we are more recently concerned is the one of Locally Decodable Codes. After having been first theoretically introduced, those codes now begin to find practical applications, most notably in cloud-based remote storage systems.

HYCOMES Team

3. Research Program

3.1. Hybrid Systems Modeling

Systems industries today make extensive use of mathematical modeling tools to design computer controlled physical systems. This class of tools addresses the modeling of physical systems with models that are simpler than usual scientific computing problems by using only Ordinary Differential Equations (ODE) and Difference Equations but not Partial Differential Equations (PDE). This family of tools first emerged in the 1980's with SystemBuild by MatrixX (now distributed by National Instruments) followed soon by Simulink by Mathworks, with an impressive subsequent development.

In the early 90's control scientists from the University of Lund (Sweden) realized that the above approach did not support component based modeling of physical systems with reuse [0]. For instance, it was not easy to draw an electrical or hydraulic circuit by assembling component models of the various devices. The development of the Omola language by Hilding Elmqvist was a first attempt to bridge this gap by supporting some form of Differential Algebraic Equations (DAE) in the models. Modelica quickly emerged from this first attempt and became in the 2000's a major international concerted effort with the Modelica Consortium [0]. A wider set of tools, both industrial and academic, now exists in this segment [0]. In the EDA sector, VHDL-AMS was developed as a standard [11].

Despite these tools are now widely used by a number of engineers, they raise a number of technical difficulties. The meaning of some programs, their mathematical semantics, can be tainted with uncertainty. A main source of difficulty lies in the failure to properly handle the discrete and the continuous parts of systems, and their interaction. How the propagation of mode changes and resets should be handled? How to avoid artifacts due to the use of a global ODE solver causing unwanted coupling between seemingly non interacting subsystems? Also, the mixed use of an equational style for the continuous dynamics with an imperative style for the mode changes and resets is a source of difficulty when handling parallel composition. It is therefore not uncommon that tools return complex warnings for programs with many different suggested hints for fixing them. Yet, these "pathological" programs can still be executed, if wanted so, giving surprising results — See for instance the Simulink examples in [6], [3] and [14].

Indeed this area suffers from the same difficulties that led to the development of the theory of synchronous languages as an effort to fix obscure compilation schemes for discrete time equation based languages in the 1980's. Our vision is that hybrid systems modeling tools deserve similar efforts in theory as synchronous languages did for the programming of embedded systems.

3.2. Background on non-standard analysis

Non-Standard analysis plays a central role in our research on hybrid systems modeling [3], [6], [15], [14]. The following text provides a brief summary of this theory and gives some hints on its usefulness in the context of hybrid systems modeling. This presentation is based on our paper [3], a chapter of Simon Bliudze's PhD thesis [21], and a recent presentation of non-standard analysis, not axiomatic in style, due to the mathematician Lindström [41].

[0] http://www.lccc.lth.se/media/LCCC2012/WorkshopSeptember/slides/Astrom.pdf
[0] https://www.modelica.org/
[0] SimScape by Mathworks, Amesim by LMS International, now Siemens PLM, and more.

Non-standard numbers allowed us to reconsider the semantics of hybrid systems and propose a radical alternative to the *super-dense time semantics* developed by Edward Lee and his team as part of the Ptolemy II project, where cascades of successive instants can occur in zero time by using $\mathbb{R}_+ \times \mathbb{N}$ as a time index. In the non-standard semantics, the time index is defined as a set $\mathbb{T} = \{n\partial \mid n \in {}^*\mathbb{N}\}$, where ∂ is an *infinitesimal* and ${}^*\mathbb{N}$ is the set of *non-standard integers*. Remark that 1/ $\mathbb{T}$ is dense in $\mathbb{R}_+$, making it "continuous", and 2/ every $t \in \mathbb{T}$ has a predecessor in $\mathbb{T}$ and a successor in $\mathbb{T}$, making it "discrete". Although it is not effective from a computability point of view, the *non-standard semantics* provides a framework that is familiar to the computer scientist and at the same time efficient as a symbolic abstraction. This makes it an excellent candidate for the development of provably correct compilation schemes and type systems for hybrid systems modeling languages.

Non-standard analysis was proposed by Abraham Robinson in the 1960s to allow the explicit manipulation of "infinitesimals" in analysis [48], [35], [10]. Robinson's approach is axiomatic; he proposes adding three new axioms to the basic Zermelo-Fraenkel (ZFC) framework. There has been much debate in the mathematical community as to whether it is worth considering non-standard analysis instead of staying with the traditional one. We do not enter this debate. The important thing for us is that non-standard analysis allows the use of the non-standard discretization of continuous dynamics "as if" it was operational.

Not surprisingly, such an idea is quite ancient. Iwasaki et al. [37] first proposed using non-standard analysis to discuss the nature of time in hybrid systems. Bliudze and Krob [22], [21] have also used non-standard analysis as a mathematical support for defining a system theory for hybrid systems. They discuss in detail the notion of "system" and investigate computability issues. The formalization they propose closely follows that of Turing machines, with a memory tape and a control mechanism.

The introduction to non-standard analysis in [21] is very pleasant and we take the liberty to borrow it. This presentation was originally due to Lindstrøm, see [41]. Its interest is that it does not require any fancy axiomatic material but only makes use of the axiom of choice — actually a weaker form of it. The proposed construction bears some resemblance to the construction of $\mathbb{R}$ as the set of equivalence classes of Cauchy sequences in $\mathbb{Q}$ modulo the equivalence relation $(u_n) \approx (v_n)$ iff $\lim_{n\to\infty}(u_n - v_n) = 0$.

3.2.1. Motivation and intuitive introduction

We begin with an intuitive introduction to the construction of the non-standard reals. The goal is to augment $\mathbb{R} \cup \{\pm\infty\}$ by adding, to each x in the set, a set of elements that are "infinitesimally close" to it. We will call the resulting set ${}^*\mathbb{R}$. Another requirement is that all operations and relations defined on $\mathbb{R}$ should extend to ${}^*\mathbb{R}$.

A first idea is to represent such additional numbers as convergent sequences of reals. For example, elements infinitesimally close to the real number zero are the sequences $u_n = 1/n$, $v_n = 1/\sqrt{n}$ and $w_n = 1/n^2$. Observe that the above three sequences can be ordered: $v_n > u_n > w_n > 0$ where 0 denotes the constant zero sequence. Of course, infinitely large elements (close to $+\infty$) can also be considered, e.g., sequences $x_u = n$, $y_n = \sqrt{n}$, and $z_n = n^2$.

Unfortunately, this way of defining ${}^*\mathbb{R}$ does not yield a total order since two sequences converging to zero cannot always be compared: if u_n and u'_n are two such sequences, the three sets $\{n \mid u_n > u'_n\}$, $\{n \mid u_n = u'_n\}$, and $\{n \mid u_n < u'_n\}$ may even all be infinitely large. The beautiful idea of Lindstrøm is to enforce that *exactly one of the above sets is important and the other two can be neglected*. This is achieved by fixing once and for all a finitely additive positive measure μ over the set $\mathbb{N}$ of integers with the following properties: [0]

1. $\mu : 2^{\mathbb{N}} \to \{0, 1\}$;
2. $\mu(X) = 0$ whenever X is finite;
3. $\mu(\mathbb{N}) = 1$.

[0] The existence of such a measure is non trivial and is explained later.

Now, once μ is fixed, one can compare any two sequences: for the above case, exactly one of the three sets must have μ-measure 1 and the others must have μ-measure 0. Thus, say that $u > u', u = u'$, or $u < u'$, if $\mu(\{n \mid u_n > u'_n\} = 1)$, $\mu(\{n \mid u_n = u'_n\}) = 1$, or $\mu(\{n \mid u_n < u'_n\}) = 1$, respectively. Indeed, the same trick works for many other relations and operations on non-standard real numbers, as we shall see. We now proceed with a more formal presentation.

3.2.2. Construction of non-standard domains

For I an arbitrary set, a *filter*$\mathcal{F}$ over I is a family of subsets of I such that:

1. the empty set does not belong to $\mathcal{F}$,
2. $P, Q \in \mathcal{F}$ implies $P \cap Q \in \mathcal{F}$, and
3. $P \in \mathcal{F}$ and $P \subset Q \subseteq I$ implies $Q \in \mathcal{F}$.

Consequently, $\mathcal{F}$ cannot contain both a set P and its complement P^c. A filter that contains one of the two for any subset $P \subseteq I$ is called an *ultra-filter*. At this point we recall Zorn's lemma, known to be equivalent to the axiom of choice:

Lemma 1 (Zorn's lemma) *Any partially ordered set $(X, \leq)$ such that any chain in X possesses an upper bound has a maximal element.*

A filter $\mathcal{F}$ over I is an ultra-filter if and only if it is maximal with respect to set inclusion. By Zorn's lemma, any filter $\mathcal{F}$ over I can be extended to an ultra-filter over I. Now, if I is infinite, the family of sets $\mathcal{F} = \{P \subseteq I \mid P^c \text{ is finite}\}$ is a *free* filter, meaning it contains no finite set. It can thus be extended to a free ultra-filter over I:

Lemma 2 Any infinite set has a free ultra-filter.

Every free ultra-filter $\mathcal{F}$ over I uniquely defines, by setting $\mu(P) = 1$ if $P \in \mathcal{F}$ and otherwise 0, a finitely additive measure [0]$\mu : 2^I \mapsto \{0, 1\}$, which satisfies

$$\mu(I) = 1 \text{ and, if } P \text{ is finite, then } \mu(P) = 0.$$

Now, fix an infinite set I and a finitely additive measure μ over I as above. Let $\mathbb{X}$ be a set and consider the Cartesian product $\mathbb{X}^I = (x_i)_{i \in I}$. Define $(x_i) \approx (x'_i)$ iff $\mu\{i \in I \mid x_i \neq x'_i\} = 0$. Relation $\approx$ is an equivalence relation whose equivalence classes are denoted by $[x_i]$ and we define:

$$^*\mathbb{X} = \mathbb{X}^I / \approx \tag{1}$$

$\mathbb{X}$ is naturally embedded into $^*\mathbb{X}$ by mapping every $x \in \mathbb{X}$ to the constant tuple such that $x_i = x$ for every $i \in I$. Any algebraic structure over $\mathbb{X}$ (group, ring, field) carries over to $^*\mathbb{X}$ by almost point-wise extension. In particular, if $[x_i] \neq 0$, meaning that $\mu\{i \mid x_i = 0\} = 0$ we can define its inverse $[x_i]^{-1}$ by taking $y_i = x_i^{-1}$ if $x_i \neq 0$ and $y_i = 0$ otherwise. This construction yields $\mu\{i \mid y_i x_i = 1\} = 1$, whence $[y_i][x_i] = 1$ in $^*\mathbb{X}$. The existence of an inverse for any non-zero element of a ring is indeed stated by the formula: $\forall x \, (x = 0 \vee \exists y \, (xy = 1))$. More generally:

Lemma 3 (Transfer Principle) Every first order formula is true over $^*\mathbb{X}$ iff it is true over $\mathbb{X}$.

The above general construction can simply be applied to $\mathbb{X} = \mathbb{R}$ and $I = \mathbb{N}$. The result is denoted $^*\mathbb{R}$; it is a field according to the transfer principle. By the same principle, $^*\mathbb{R}$ is totally ordered by $[u_n] \leq [v_n]$ iff $\mu\{n \mid u_n > v_n\} = 0$. We claim that, for any finite $[x_n] \in {}^*\mathbb{R}$, there exists a unique $st([x_n])$, call it the *standard part* of $[x_n]$, such that

$$st([x_n]) \in \mathbb{R} \text{ and } st([x_n]) \approx [x_n]\,. \tag{2}$$

[0]Observe that, as a consequence, μ cannot be sigma-additive (in contrast to probability measures or Radon measures) in that it is *not* true that $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$ holds for an infinite denumerable sequence A_n of pairwise disjoint subsets of $\mathbb{N}$.

To prove this, let $x = \sup\{u \in \mathbb{R} \mid [u] \leq [x_n]\}$, where $[u]$ denotes the constant sequence equal to u. Since $[x_n]$ is finite, x exists and we only need to show that $[x_n] - x$ is infinitesimal. If not, then there exists $y \in \mathbb{R}, y > 0$ such that $y < |x - [x_n]|$, that is, either $x < [x_n] - [y]$ or $x > [x_n] + [y]$, which both contradict the definition of x. The uniqueness of x is clear, thus we can define $st([x_n]) = x$. Infinite non-standard reals have no standard part in $\mathbb{R}$.

It is also of interest to apply the general construction (1) to $\mathbb{X} = I = \mathbb{N}$, which results in the set ${}^*\mathbb{N}$ of *non-standard natural numbers*. The non-standard set ${}^*\mathbb{N}$ differs from $\mathbb{N}$ by the addition of *infinite natural numbers*, which are equivalence classes of sequences of integers whose essential limit is $+\infty$.

3.3. Contract-Based Design, Interfaces Theories, and Requirements Engineering

System companies such as automotive and aeronautic companies are facing significant difficulties due to the exponentially raising complexity of their products coupled with increasingly tight demands on functionality, correctness, and time-to-market. The cost of being late to market or of imperfections in the products is staggering as witnessed by the recent recalls and delivery delays that many major car and airplane manufacturers had to bear in the recent years. The specific root causes of these design problems are complex and relate to a number of issues ranging from design processes and relationships with different departments of the same company and with suppliers, to incomplete requirement specification and testing.

We believe the most promising means to address the challenges in systems engineering is to employ structured and formal design methodologies that seamlessly and coherently combine the various viewpoints of the design space (behavior, space, time, energy, reliability, ...), that provide the appropriate abstractions to manage the inherent complexity, and that can provide correct-by-construction implementations. The following technology issues must be addressed when developing new approaches to the design of complex systems:

- The overall design flows for heterogeneous systems and the associated use of models across traditional boundaries are not well developed and understood. Relationships between different teams inside a same company, or between different stake-holders in the supplier chain, are not well supported by solid technical descriptions for the mutual obligations.
- System requirements capture and analysis is in large part a heuristic process, where the informal text and natural language-based techniques in use today are facing significant challenges. Formal requirements engineering is in its infancy: mathematical models, formal analysis techniques and links to system implementation must be developed.
- Dealing with variability, uncertainty, and life-cycle issues, such as extensibility of a product family, are not well-addressed using available systems engineering methodologies and tools.

The challenge is to address the entire process and not to consider only local solutions of methodology, tools, and models that ease part of the design.

Contract-based design has been proposed as a new approach to the system design problem that is rigorous and effective in dealing with the problems and challenges described before, and that, at the same time, does not require a radical change in the way industrial designers carry out their task as it cuts across design flows of different type. Indeed, contracts can be used almost everywhere and at nearly all stages of system design, from early requirements capture, to embedded computing infrastructure and detailed design involving circuits and other hardware. Contracts explicitly handle pairs of properties, respectively representing the assumptions on the environment and the guarantees of the system under these assumptions. Intuitively, a contract is a pair $C = (A, G)$ of assumptions and guarantees characterizing in a formal way 1) under which context the design is assumed to operate, and 2) what its obligations are. Assume/Guarantee reasoning has been known for a long time, and has been used mostly as verification mean for the design of software [45]. However, contract based design with explicit assumptions is a philosophy that should be followed all along the design, with all kinds of models, whenever necessary. Here, specifications are not limited to profiles, types, or taxonomy of data, but also describe the functions, performances of various kinds (time and energy), and reliability. This amounts to enrich a component's interface with, on one hand, formal specifications of the behavior of the environment in

which the component may be instantiated and, on the other hand, of the expected behavior of the component itself. The consideration of rich interfaces is still in its infancy. So far, academic researchers have addressed the mathematics and algorithmics of interfaces theories and contract-based reasoning. To make them a technique of choice for system engineers, we must develop:

- Mathematical foundations for interfaces and requirements engineering that enable the design of frameworks and tools;
- A system engineering framework and associated methodologies and tool sets that focus on system requirements modeling, contract specification, and verification at multiple abstraction layers.

A detailed bibliography on contract and interface theories for embedded system design can be found in [4]. In a nutshell, contract and interface theories fall into two main categories:

Assume/guarantee contracts. By explicitly relying on the notions of assumptions and guarantees, A/G-contracts are intuitive, which makes them appealing for the engineer. In A/G-contracts, assumptions and guarantees are just properties regarding the behavior of a component and of its environment. The typical case is when these properties are formal languages or sets of traces, which includes the class of safety properties [38], [29], [44], [13], [30]. Contract theories were initially developed as specification formalisms able to refuse some inputs from the environment [36]. A/G-contracts were advocated by the SPEEDS project [16]. They were further experimented in the framework of the CESAR project [31], with the additional consideration of *weak* and *strong* assumptions. This is still a very active research topic, with several recent contributions dealing with the timed [20] and probabilistic [25], [26] viewpoints in system design, and even mixed-analog circuit design [46].

Automata theoretic interfaces. Interfaces combine assumptions and guarantees in a single, automata theoretic specification. Most interface theories are based on Lynch Input/Output Automata [43], [42]. Interface Automata [51], [50], [52], [27] focus primarily on parallel composition and compatibility: Two interfaces can be composed and are compatible if there is at least one environment where they can work together. The idea is that the resulting composition exposes as an interface the needed information to ensure that incompatible pairs of states cannot be reached. This can be achieved by using the possibility, for an Interface Automaton, to refuse selected inputs from the environment in a given state, which amounts to the implicit assumption that the environment will never produce any of the refused inputs, when the interface is in this state. Modal Interfaces [5] inherit from both Interface Automata and the originally unrelated notion of Modal Transition System [40], [12], [23], [39]. Modal Interfaces are strictly more expressive than Interface Automata by decoupling the I/O orientation of an event and its deontic modalities (mandatory, allowed or forbidden). Informally, a *must* transition is available in every component that realizes the modal interface, while a *may* transition needs not be. Research on interface theories is still very active. For instance, timed [53], [17], [19], [33], [32], [18], probabilistic [25], [34] and energy-aware [28] interface theories have been proposed recently.

Requirements Engineering is one of the major concerns in large systems industries today, particularly so in sectors where certification prevails [49]. DOORS projects collecting requirements are poorly structured and cannot be considered a formal modeling framework today. They are nothing more than an informal documentation enriched with hyperlinks. As examples, medium size sub-systems may have a few thousands requirements and the Rafale fighter aircraft has above 250,000 of them. For the Boeing 787, requirements were not stable while subcontractors performed the development of the fly-by-wire and of the landing gear subsystems.

We see Contract-Based Design and Interfaces Theories as innovative tools in support of Requirements Engineering. The Software Engineering community has extensively covered several aspects of Requirements Engineering, in particular:

- the development and use of large and rich *ontologies*; and
- the use of Model Driven Engineering technology for the structural aspects of requirements and resulting hyperlinks (to tests, documentation, PLM, architecture, and so on).

Behavioral models and properties, however, are not properly encompassed by the above approaches. This is the cause of a remaining gap between this phase of systems design and later phases where formal model based methods involving behavior have become prevalent—see the success of Matlab/Simulink/Scade technologies. We believe that our work on contract based design and interface theories is best suited to bridge this gap.

LFANT Project-Team

3. Research Program

3.1. Number fields, class groups and other invariants

Participants: Bill Allombert, Athanasios Angelakis, Karim Belabas, Julio Brau Avila, Jean-Paul Cerri, Henri Cohen, Jean-Marc Couveignes, Andreas Enge, Pınar Kılıçer, Pierre Lezowski, Nicolas Mascot, Aurel Page.

Modern number theory has been introduced in the second half of the 19th century by Dedekind, Kummer, Kronecker, Weber and others, motivated by Fermat's conjecture: There is no non-trivial solution in integers to the equation $x^n + y^n = z^n$ for $n \geqslant 3$. For recent textbooks, see [5]. Kummer's idea for solving Fermat's problem was to rewrite the equation as $(x+y)(x+\zeta y)(x+\zeta^2 y)\cdots(x+\zeta^{n-1}y) = z^n$ for a primitive n-th root of unity ζ, which seems to imply that each factor on the left hand side is an n-th power, from which a contradiction can be derived.

The solution requires to augment the integers by *algebraic numbers*, that are roots of polynomials in $\mathbb{Z}[X]$. For instance, ζ is a root of $X^n - 1$, $\sqrt[3]{2}$ is a root of $X^3 - 2$ and $\frac{\sqrt{3}}{5}$ is a root of $25X^2 - 3$. A *number field* consists of the rationals to which have been added finitely many algebraic numbers together with their sums, differences, products and quotients. It turns out that actually one generator suffices, and any number field K is isomorphic to $\mathbb{Q}[X]/(f(X))$, where $f(X)$ is the minimal polynomial of the generator. Of special interest are *algebraic integers*, "numbers without denominators", that are roots of a monic polynomial. For instance, ζ and $\sqrt[3]{2}$ are integers, while $\frac{\sqrt{3}}{5}$ is not. The *ring of integers* of K is denoted by $\mathcal{O}_K$; it plays the same role in K as $\mathbb{Z}$ in $\mathbb{Q}$.

Unfortunately, elements in $\mathcal{O}_K$ may factor in different ways, which invalidates Kummer's argumentation. Unique factorisation may be recovered by switching to *ideals*, subsets of $\mathcal{O}_K$ that are closed under addition and under multiplication by elements of $\mathcal{O}_K$. In $\mathbb{Z}$, for instance, any ideal is *principal*, that is, generated by one element, so that ideals and numbers are essentially the same. In particular, the unique factorisation of ideals then implies the unique factorisation of numbers. In general, this is not the case, and the *class group* Cl_K of ideals of $\mathcal{O}_K$ modulo principal ideals and its *class number* $h_K = |\mathrm{Cl}_K|$ measure how far $\mathcal{O}_K$ is from behaving like $\mathbb{Z}$.

Using ideals introduces the additional difficulty of having to deal with *units*, the invertible elements of $\mathcal{O}_K$: Even when $h_K = 1$, a factorisation of ideals does not immediately yield a factorisation of numbers, since ideal generators are only defined up to units. For instance, the ideal factorisation $(6) = (2)\cdot(3)$ corresponds to the two factorisations $6 = 2\cdot 3$ and $6 = (-2)\cdot(-3)$. While in $\mathbb{Z}$, the only units are 1 and -1, the unit structure in general is that of a finitely generated $\mathbb{Z}$-module, whose generators are the *fundamental units*. The *regulator* R_K measures the "size" of the fundamental units as the volume of an associated lattice.

One of the main concerns of algorithmic algebraic number theory is to explicitly compute these invariants (Cl_K and h_K, fundamental units and R_K), as well as to provide the data allowing to efficiently compute with numbers and ideals of $\mathcal{O}_K$; see [35] for a recent account.

The *analytic class number formula* links the invariants h_K and R_K (unfortunately, only their product) to the ζ-function of K, $\zeta_K(s) := \prod_{\mathfrak{p} \text{ prime ideal of } \mathcal{O}_K} (1 - \mathrm{N}\,\mathfrak{p}^{-s})^{-1}$, which is meaningful when $\Re(s) > 1$, but which may be extended to arbitrary complex $s \neq 1$. Introducing characters on the class group yields a generalisation of ζ- to L-functions. The *generalised Riemann hypothesis (GRH)*, which remains unproved even over the rationals, states that any such L-function does not vanish in the right half-plane $\Re(s) > 1/2$. The validity of the GRH has a dramatic impact on the performance of number theoretic algorithms. For instance, under GRH, the class group admits a system of generators of polynomial size; without GRH, only exponential bounds are known. Consequently, an algorithm to compute Cl_K via generators and relations (currently the only viable practical approach) either has to assume that GRH is true or immediately becomes exponential.

When $h_K = 1$ the number field K may be norm-Euclidean, endowing $\mathcal{O}_K$ with a Euclidean division algorithm. This question leads to the notions of the Euclidean minimum and spectrum of K, and another task in algorithmic number theory is to compute explicitly this minimum and the upper part of this spectrum, yielding for instance generalised Euclidean gcd algorithms.

3.2. Function fields, algebraic curves and cryptology

Participants: Karim Belabas, Julio Brau Avila, Jean-Marc Couveignes, Andreas Enge, Hamish Ivey-Law, Nicolas Mascot, Enea Milio, Damien Robert.

Algebraic curves over finite fields are used to build the currently most competitive public key cryptosystems. Such a curve is given by a bivariate equation $\mathcal{C}(X, Y) = 0$ with coefficients in a finite field $\mathbb{F}_q$. The main classes of curves that are interesting from a cryptographic perspective are *elliptic curves* of equation $\mathcal{C} = Y^2 - (X^3 + aX + b)$ and *hyperelliptic curves* of equation $\mathcal{C} = Y^2 - (X^{2g+1} + \cdots)$ with $g \geqslant 2$.

The cryptosystem is implemented in an associated finite abelian group, the *Jacobian* $\mathrm{Jac}_\mathcal{C}$. Using the language of function fields exhibits a close analogy to the number fields discussed in the previous section. Let $\mathbb{F}_q(X)$ (the analogue of $\mathbb{Q}$) be the *rational function field* with subring $\mathbb{F}_q[X]$ (which is principal just as $\mathbb{Z}$). The *function field* of $\mathcal{C}$ is $K_\mathcal{C} = \mathbb{F}_q(X)[Y]/(\mathcal{C})$; it contains the *coordinate ring* $\mathcal{O}_\mathcal{C} = \mathbb{F}_q[X, Y]/(\mathcal{C})$. Definitions and properties carry over from the number field case $K/\mathbb{Q}$ to the function field extension $K_\mathcal{C}/\mathbb{F}_q(X)$. The Jacobian $\mathrm{Jac}_\mathcal{C}$ is the divisor class group of $K_\mathcal{C}$, which is an extension of (and for the curves used in cryptography usually equals) the ideal class group of $\mathcal{O}_\mathcal{C}$.

The size of the Jacobian group, the main security parameter of the cryptosystem, is given by an L-function. The GRH for function fields, which has been proved by Weil, yields the Hasse–Weil bound $(\sqrt{q} - 1)^{2g} \leqslant |\mathrm{Jac}_\mathcal{C}| \leqslant (\sqrt{q} + 1)^{2g}$, or $|\mathrm{Jac}_\mathcal{C}| \approx q^g$, where the *genus* g is an invariant of the curve that correlates with the degree of its equation. For instance, the genus of an elliptic curve is 1, that of a hyperelliptic one is $\frac{\deg_X \mathcal{C} - 1}{2}$. An important algorithmic question is to compute the exact cardinality of the Jacobian.

The security of the cryptosystem requires more precisely that the *discrete logarithm problem* (DLP) be difficult in the underlying group; that is, given elements D_1 and $D_2 = xD_1$ of $\mathrm{Jac}_\mathcal{C}$, it must be difficult to determine x. Computing x corresponds in fact to computing $\mathrm{Jac}_\mathcal{C}$ explicitly with an isomorphism to an abstract product of finite cyclic groups; in this sense, the DLP amounts to computing the class group in the function field setting.

For any integer n, the *Weil pairing* e_n on $\mathcal{C}$ is a function that takes as input two elements of order n of $\mathrm{Jac}_\mathcal{C}$ and maps them into the multiplicative group of a finite field extension $\mathbb{F}_{q^k}$ with $k = k(n)$ depending on n. It is bilinear in both its arguments, which allows to transport the DLP from a curve into a finite field, where it is potentially easier to solve. The *Tate-Lichtenbaum pairing*, that is more difficult to define, but more efficient to implement, has similar properties. From a constructive point of view, the last few years have seen a wealth of cryptosystems with attractive novel properties relying on pairings.

For a random curve, the parameter k usually becomes so big that the result of a pairing cannot even be output any more. One of the major algorithmic problems related to pairings is thus the construction of curves with a given, smallish k.

3.3. Complex multiplication

Participants: Karim Belabas, Henri Cohen, Jean-Marc Couveignes, Andreas Enge, Hamish Ivey-Law, Chloë Martindale, Nicolas Mascot, Enea Milio, Aurel Page, Damien Robert.

Complex multiplication provides a link between number fields and algebraic curves; for a concise introduction in the elliptic curve case, see [37], for more background material, [36]. In fact, for most curves $\mathcal{C}$ over a finite field, the endomorphism ring of $\mathrm{Jac}_\mathcal{C}$, which determines its L-function and thus its cardinality, is an order in a special kind of number field K, called *CM field*. The CM field of an elliptic curve is an imaginary-quadratic field $\mathbb{Q}(\sqrt{D})$ with $D < 0$, that of a hyperelliptic curve of genus g is an imaginary-quadratic extension of a totally real number field of degree g. Deuring's lifting theorem ensures that $\mathcal{C}$ is the reduction modulo some prime of a curve with the same endomorphism ring, but defined over the *Hilbert class field* H_K of K.

Algebraically, H_K is defined as the maximal unramified abelian extension of K; the Galois group of H_K/K is then precisely the class group Cl_K. A number field extension H/K is called *Galois* if $H \simeq K[X]/(f)$ and H contains all complex roots of f. For instance, $\mathbb{Q}(\sqrt{2})$ is Galois since it contains not only $\sqrt{2}$, but also the second root $-\sqrt{2}$ of X^2-2, whereas $\mathbb{Q}(\sqrt[3]{2})$ is not Galois, since it does not contain the root $e^{2\pi i/3}\sqrt[3]{2}$ of X^3-2. The *Galois group* $\mathrm{Gal}_{H/K}$ is the group of automorphisms of H that fix K; it permutes the roots of f. Finally, an *abelian* extension is a Galois extension with abelian Galois group.

Analytically, in the elliptic case H_K may be obtained by adjoining to K the *singular value* $j(\tau)$ for a complex valued, so-called *modular* function j in some $\tau \in \mathcal{O}_K$; the correspondence between $\mathrm{Gal}_{H/K}$ and Cl_K allows to obtain the different roots of the minimal polynomial f of $j(\tau)$ and finally f itself. A similar, more involved construction can be used for hyperelliptic curves. This direct application of complex multiplication yields algebraic curves whose L-functions are known beforehand; in particular, it is the only possible way of obtaining ordinary curves for pairing-based cryptosystems.

The same theory can be used to develop algorithms that, given an arbitrary curve over a finite field, compute its L-function.

A generalisation is provided by *ray class fields*; these are still abelian, but allow for some well-controlled ramification. The tools for explicitly constructing such class fields are similar to those used for Hilbert class fields.

MARELLE Project-Team

3. Research Program

3.1. Type theory and formalization of mathematics

The calculus of inductive constructions is a branch of type theory that serves as a foundation for theorem proving tools, especially the Coq proof assistant. It is powerful enough to formalize complex mathematics, based on algebraic structures and operations. This is especially important as we want to produce proofs of logical properties for these algebraic structures, a goal that is only marginally addressed in most scientific computation systems.

The calculus of inductive constructions also makes it possible to write algorithms as recursive functional programs which manipulate tree-like data structures. A third important characteristic of this calculus is that it is also a language for manipulating proofs. All this makes this calculus a tool of choice for our investigations. However, this language is still being improved and part of our work concerns these improvements.

3.2. Verification of scientific algorithms

To produce certified algorithms, we use the following approach: instead of attempting to prove properties of an existing program written in a conventional programming language such as C or Java, we produce new programs in the calculus of constructions whose correctness is an immediate consequence of their construction. This has several advantages. First, we work at a high level of abstraction, independently of the target implementation language. Secondly, we concentrate on specific characteristics of the algorithm, and abstract away from the rest (for instance, we abstract away from memory management or data implementation strategies). Therefore, we are able to address more high-level mathematics and to express more general properties without being overwhelmed by implementation details.

However, this approach also presents a few drawbacks. For instance, the calculus of constructions usually imposes that recursive programs should explicitly terminate for all inputs. For some algorithms, we need to use advanced concepts (for instance, well-founded relations) to make the property of termination explicit, and proofs of correctness become especially difficult in this setting.

3.3. Programming language semantics

To bridge the gap between our high-level descriptions of algorithms and conventional programming languages, we investigate the algorithms that are present in programming language implementations, for instance algorithms that are used in a compiler or a static analysis tool. For these algorithms, we generally base our work on the semantic description of a language. The properties that we attempt to prove for an algorithm are, for example, that an optimization respects the meaning of programs or that the programs produced are free of some unwanted behavior. In practice, we rely on this study of programming language semantics to propose extensions to theorem proving tools or to participate in the verification that compilers for conventional programming languages are exempt from bugs.

MEXICO Project-Team

3. Research Program

3.1. Concurrency

Participants: Benedikt Bollig, Thomas Chatain, Aiswarya Cyriac, Paul Gastin, Stefan Haar, Serge Haddad, Hernán Ponce de Léon, Stefan Schwoon.

Concurrency: Property of systems allowing some interacting processes to be executed in parallel.

Diagnosis: The process of deducing from a partial observation of a system aspects of the internal states or events of that system; in particular, *fault diagnosis* aims at determining whether or not some non-observable fault event has occurred.

Conformance Testing: Feeding dedicated input into an implemented system IS and deducing, from the resulting output of I, whether I respects a formal specification S.

3.1.1. Introduction

It is well known that, whatever the intended form of analysis or control, a *global* view of the system state leads to overwhelming numbers of states and transitions, thus slowing down algorithms that need to explore the state space. Worse yet, it often blurs the mechanics that are at work rather than exhibiting them. Conversely, respecting concurrency relations avoids exhaustive enumeration of interleavings. It allows us to focus on 'essential' properties of non-sequential processes, which are expressible with causal precedence relations. These precedence relations are usually called causal (partial) orders. Concurrency is the explicit absence of such a precedence between actions that do not have to wait for one another. Both causal orders and concurrency are in fact essential elements of a specification. This is especially true when the specification is constructed in a distributed and modular way. Making these ordering relations explicit requires to leave the framework of state/interleaving based semantics. Therefore, we need to develop new dedicated algorithms for tasks such as conformance testing, fault diagnosis, or control for distributed discrete systems. Existing solutions for these problems often rely on centralized sequential models which do not scale up well.

3.1.2. Diagnosis

Participants: Benedikt Bollig, Stefan Haar, Serge Haddad, Loig Jezequel, Hernán Ponce de Léon, Stefan Schwoon.

Fault Diagnosis for discrete event systems is a crucial task in automatic control. Our focus is on *event oriented* (as opposed to *state oriented*) model-based diagnosis, asking e.g. the following questions:
given a - potentially large - *alarm pattern* formed of observations,

- what are the possible *fault scenarios* in the system that *explain* the pattern ?
- Based on the observations, can we deduce whether or not a certain - invisible - fault has actually occurred ?

Model-based diagnosis starts from a discrete event model of the observed system - or rather, its relevant aspects, such as possible fault propagations, abstracting away other dimensions. From this model, an extraction or unfolding process, guided by the observation, produces recursively the explanation candidates.

In asynchronous partial-order based diagnosis with Petri nets [63], [64], [68], one unfolds the *labelled product* of a Petri net model $\mathcal{N}$ and an observed alarm pattern $\mathcal{A}$, also in Petri net form. We obtain an acyclic net giving partial order representation of the behaviors compatible with the alarm pattern. A recursive online procedure filters out those runs *(configurations)* that explain *exactly*$\mathcal{A}$. The Petri-net based approach generalizes to dynamically evolving topologies, in dynamical systems modeled by graph grammars, see [47]

3.1.2.1. Observability and Diagnosability

Diagnosis algorithms have to operate in contexts with low observability, i.e., in systems where many events are invisible to the supervisor. Checking *observability* and *diagnosability* for the supervised systems is therefore a crucial and non-trivial task in its own right. Analysis of the relational structure of occurrence nets allows us to check whether the system exhibits sufficient visibility to allow diagnosis. Developing efficient methods for both verification of *diagnosability checking* under concurrency, and the *diagnosis* itself for distributed, composite and asynchronous systems, is an important field for *MExICo*.

3.1.2.2. Distribution

Distributed computation of unfoldings allows one to factor the unfolding of the global system into smaller *local* unfoldings, by local supervisors associated with sub-networks and communicating among each other. In [64], [49], elements of a methodology for distributed computation of unfoldings between several supervisors, underwritten by algebraic properties of the category of Petri nets have been developed. Generalizations, in particular to Graph Grammars, are still do be done.

Computing diagnosis in a distributed way is only one aspect of a much vaster topic, that of *distributed diagnosis* (see [60], [73]). In fact, it involves a more abstract and often indirect reasoning to conclude whether or not some given invisible fault has occurred. Combination of local scenarios is in general not sufficient: the global system may have behaviors that do not reveal themselves as faulty (or, dually, non-faulty) on any local supervisor's domain (compare [46], [52]). Rather, the local diagnosers have to join all *information* that is available to them locally, and then deduce collectively further information from the combination of their views. In particular, even the *absence* of fault evidence on all peers may allow to deduce fault occurrence jointly, see [78], [79]. Automatizing such procedures for the supervision and management of distributed and locally monitored asynchronous systems is a long-term goal to which *MExICo* hopes to contribute.

3.1.3. Contextual nets

Participant: Stefan Schwoon.

Assuring the correctness of concurrent systems is notoriously difficult due to the many unforeseeable ways in which the components may interact and the resulting state-space explosion. A well-established approach to alleviate this problem is to model concurrent systems as Petri nets and analyse their unfoldings, essentially an acyclic version of the Petri net whose simpler structure permits easier analysis [62].

However, Petri nets are inadequate to model concurrent read accesses to the same resource. Such situations often arise naturally, for instance in concurrent databases or in asynchronous circuits. The encoding tricks typically used to model these cases in Petri nets make the unfolding technique inefficient. Contextual nets, which explicitly do model concurrent read accesses, address this problem. Their accurate representation of concurrency makes contextual unfoldings up to exponentially smaller in certain situations. An abstract algorithm for contextual unfoldings was first given in [48]. In recent work, we further studied this subject from a theoretical and practical perspective, allowing us to develop concrete, efficient data structures and algorithms and a tool (Cunf) that improves upon existing state of the art. This work led to the PhD thesis of César Rodríguez.

Contexutal unfoldings deal well with two sources of state-space explosion: concurrency and shared resources. Recently, we proposed an improved data structure, called *contextual merged processes* (CMP) to deal with a third source of state-space explosion, i.e. sequences of choices. The work on CMP [81] is currently at an abstract level. In the short term, we want to put this work into practice, requiring some theoretical groundwork, as well as programming and experimentation.

Another well-known approach to verifying concurrent systems is *partial-order reduction*, exemplified by the tool SPIN. Although it is known that both partial-order reduction and unfoldings have their respective strengths and weaknesses, we are not aware of any conclusive comparison between the two techniques. Spin comes with a high-level modeling language having an explicit notion of processes, communication channels, and variables. Indeed, the reduction techniques implemented in Spin exploit the specific properties of these features. On the other side, while there exist highly efficient tools for unfoldings, Petri nets are a relatively general low-level

formalism, so these techniques do not exploit properties of higher language features. Our work on contextual unfoldings and CMPs represents a first step to make unfoldings exploit richer models. In the long run, we wish raise the unfolding technique to a suitable high-level modelling language and develop appropriate tool support.

3.1.4. Verification of Concurrent Recursive Programs

Participants: Benedikt Bollig, Aiswarya Cyriac, Paul Gastin, Stefan Schwoon.

In a DIGITEO PhD project, we will study logical specification formalisms for concurrent recursive programs. With the advent of multi-core processors, the analysis and synthesis of such programs is becoming more and more important. However, it cannot be achieved without more comprehensive formal mathematical models of concurrency and parallelization. Most existing approaches have in common that they restrict to the analysis of an over- or underapproximation of the actual program executions and do not focus on a behavioral semantics. In particular, temporal logics have not been considered. Their design and study will require the combination of prior works on logics for sequential recursive programs and concurrent finite-state programs.

3.1.5. Dynamic and parameterized concurrent systems

Participants: Benedikt Bollig, Paul Gastin.

In the past few years, our research has focused on concurrent systems where the architecture, which provides a set of processes and links between them, is *static* and *fixed in advance*. However, the assumption that the set of processes is fixed somehow seems to hinder the application of formal methods in practice. It is not appropriate in areas such as mobile computing or ad-hoc networks. In concurrent programming, it is actually perfectly natural to design a program, and claim its correctness, independently of the number of processes that participate in its execution. There are, essentially, two kinds of systems that fall into this category. When the process architecture is static but unknown, it is a parameter of the system; we then call a system *parameterized*. When, on the other hand, the process architecure is generated at runtime (i.e., process creation is a communication primitive), we say that a system is *dynamic*. Though parameterized and dynamic systems have received increasing interest in recent years, there is, by now, no canonical approach to modeling and verifying such systems. Our research program aims at the development of *a theory of parameterized and dynamic concurrent systems.* More precisely, our goal is a *unifying* theory that lays algebraic, logical, and automata-theoretic foundations to support and facilitate the study of parameterized and dynamic concurrent systems. Such theories indeed exist in non-parameterized settings where the number of processes and the way they are connected are fixed in advance. However, parameterized and dynamic systems lack such foundations and often restict to very particular models with specialized verification techniques.

3.1.6. Testing

Participants: Benedikt Bollig, Paul Gastin, Stefan Haar, Hernán Ponce de Léon.

3.1.6.1. Introduction

The gap between specification and implementation is at the heart of research on formal testing. The general *conformance testing problem* can be defined as follows: Does an implementation $\mathcal{M}'$ conform a given specification $\mathcal{M}$? Here, both $\mathcal{M}$ and $\mathcal{M}'$ are assumed to have input and output channels. The formal model $\mathcal{M}$ of the specification is entirely known and can be used for analysis. On the other hand, the implementation $\mathcal{M}'$ is unknown but interacts with the environment through observable input and output channels. So the behavior of $\mathcal{M}'$ is partially controlled by input streams, and partially observable via output streams. The Testing problem consists in computing, from the knowledge of $\mathcal{M}$, *input streams* for $\mathcal{M}'$ such that observation of the resulting output streams from $\mathcal{M}'$ allows to determine whether $\mathcal{M}'$ conforms to $\mathcal{M}$ as intended.

In this project, we focus on distributed or asynchronous versions of the conformance testing problem. There are two main difficulties. First, due to the distributed nature of the system, it may not be possible to have a unique global observer for the outcome of a test. Hence, we may need to use *local* observers which will record only *partial views* of the execution. Due to this, it is difficult or even impossible to reconstruct a coherent global execution. The second difficulty is the lack of global synchronization in distributed asynchronous systems. Up to now, models were described with I/O automata having a centralized control, hence inducing global synchronizations.

3.1.6.2. Asynchronous Testing

Since 2006 and in particular during his sabbatical stay at the University of Ottawa, Stefan Haar has been working with Guy-Vincent Jourdan and Gregor v. Bochmann of UOttawa and Claude Jard of IRISA on asynchronous testing. In the synchronous (sequential) approach, the model is described by an I/O automaton with a centralized control and transitions labeled with individual input or output actions. This approach has known limitations when inputs and outputs are distributed over remote sites, a feature that is characteristic of , e.g., web computing. To account for concurrency in the system, they have developed in [70], [53] asynchronous conformance testing for automata with transitions labeled with (finite) partial orders of I/O. Intuitively, this is a "big step" semantics where each step allows concurrency but the system is synchronized before the next big step. This is already an important improvement on the synchronous setting. The non-trivial challenge is now to cope with fully asynchronous specifications using models with decentralized control such as Petri nets.

3.1.6.3. Near Future

Completion of asynchronous testing in the setting without any big-step synchronization, and an improved understanding of the relations and possible interconnections between local (i.e. distributed) and asynchronous (centralized) testing. This has been the objective of the *TECSTES* project (2011-2014), funded by a DIGITEO *DIM/LSC* grant, and which involved Hernán Ponce de Léon and Stefan Haar of *MExICo*, and Delphine Longuet at LRI, University Paris-Sud/Orsay. We have extended several well known conformance (ioco style) relations for sequential models to models that can handle concurrency (labeled event structures). Two semantics (interleaving and partial order) were presented for every relation. With the interleaving semantics, the relations we obtained boil down to the same relations defined for labeled transition systems, since they focus on sequences of actions. The only advantage of using labeled event structures as a specification formalism for testing remains in the conciseness of the concurrent model with respect to a sequential one. As far as testing is concerned, the benefit is low since every interleaving has to be tested. By contrast, under the partial order semantics, the relations we obtain allow to distinguish explicitly implementations where concurrent actions are implemented concurrently, from those where they are interleaved, i.e. implemented sequentially. Therefore, these relations will be of interest when designing distributed systems, since the natural concurrency between actions that are performed in parallel by different processes can be taken into account. In particular, the fact of being unable to control or observe the order between actions taking place on different processes will not be considered as an impediment for testing. We have developped a complete testing framework for concurrent systems, which included the notions of test suites and test cases. We studied what kind of systems are testable in such a framework, and we have proposed sufficient conditions for obtaining a complete test suite as well as an algorithm to construct a test suite with such properties.

A mid-to long term goal (which may or may not be addressed by *MExICo* depending on the availability of staff for this subject) is the comprehensive formalization of testing and testability in asynchronous systems with distributed architecture and test protocols.

3.2. Interaction

Participants: Benedikt Bollig, Thomas Chatain, Paul Gastin, Stefan Haar, Serge Haddad.

3.2.1. Introduction

Systems and services exhibit non-trivial *interaction* between specialized and heterogeneous components. This interplay is challenging for several reasons. On one hand, a coordinated interplay of several components is required, though each has only a limited, partial view of the system's configuration. We refer to this problem as *distributed synthesis* or *distributed control*. An aggravating factor is that the structure of a component might be semi-transparent, which requires a form of *grey box management*.

Interaction, one of the main characteristics of systems under consideration, often involves an environment that is not under the control of cooperating services. To achieve a common goal, the services need to agree upon a strategy that allows them to react appropriately regardless of the interactions with the environment. Clearly, the notions of opponents and strategies fall within *game theory*, which is naturally one of our main tools in exploring interaction. We will apply to our problems techniques and results developed in the domains

of distributed games and of games with partial information. We will consider also new problems on games that arise from our applications.

3.2.2. Distributed Control

Participants: Benedikt Bollig, Thomas Chatain, Paul Gastin, Stefan Haar.

Program synthesis, as introduced by Church [59] aims at deriving directly an implementation from a specification, allowing the implementation to be correct by design. When the implementation is already at hand but choices remain to be resolved at run time then the problem becomes controller synthesis. Both program and controller synthesis have been extensively studied for sequential systems. In a distributed setting, we need to synthesize a distributed program or distributed controllers that interact locally with the system components. The main difficulty comes from the fact that the local controllers/programs have only a partial view of the entire system. This is also an old problem largely considered undecidable in most settings [77], [72], [75], [65], [67].

Actually, the main undecidability sources come from the fact that this problem was addressed in a synchronous setting using global runs viewed as sequences. In a truly distributed system where interactions are asynchronous we have recently obtained encouraging decidability results [66], [57]. This is a clear witness where concurrency may be exploited to obtain positive results. It is essential to specify expected properties directly in terms of causality revealed by partial order models of executions (MSCs or Mazurkiewicz traces). We intend to develop this line of research with the ambitious aim to obtain decidability for all natural systems and specifications. More precisely, we will identify natural hypotheses both on the architecture of our distributed system and on the specifications under which the distributed program/controller synthesis problem is decidable. This should open the way to important applications, e.g., for distributed control of embedded systems.

3.2.3. Adaptation and Grey box management

Participants: Stefan Haar, Serge Haddad.

Contrary to mainframe systems or monolithic applications of the past, we are experiencing and using an increasing number of services that are performed not by one provider but rather by the interaction and cooperation of many specialized components. As these components come from different providers, one can no longer assume all of their internal technologies to be known (as it is the case with proprietary technology). Thus, in order to compose e.g. orchestrated services over the web, to determine violations of specifications or contracts, to adapt existing services to new situations etc, one needs to analyze the interaction behavior of *boxes* that are known only through their public interfaces. For their semi-transparent-semi-opaque nature, we shall refer to them as **grey boxes**. While the concrete nature of these boxes can range from vehicles in a highway section to hotel reservation systems, the tasks of *grey box management* have universal features allowing for generalized approaches with formal methods. Two central issues emerge:

- Abstraction: From the designer point of view, there is a need for a trade-off between transparency (no abstraction) in order to integrate the box in different contexts and opacity (full abstraction) for security reasons.
- Adaptation: Since a grey box gives a partial view about the behavior of the component, even if it is not immediately useable in some context, the design of an adaptator is possible. Thus the goal is the synthesis of such an adaptator from a formal specification of the component and the environment.

Our work on direct modeling and handling of "grey boxes" via modal models (see [61]) was halted when Dorsaf El-Hog stopped her PhD work to leave academia, and has not resumed for lack of staff. However, it should be noted that semi-transparent system management in a larger sense remains an active field for the team, witness in particular our work on diagnosis and testing.

3.3. Management of Quantitative Behavior

Participants: Benedikt Bollig, Thomas Chatain, Paul Gastin, Stefan Haar, Serge Haddad, Benjamin Monmege.

3.3.1. Introduction

Besides the logical functionalities of programs, the *quantitative* aspects of component behavior and interaction play an increasingly important role.

- *Real-time* properties cannot be neglected even if time is not an explicit functional issue, since transmission delays, parallelism, etc, can lead to time-outs striking, and thus change even the logical course of processes. Again, this phenomenon arises in telecommunications and web services, but also in transport systems.
- In the same contexts, *probabilities* need to be taken into account, for many diverse reasons such as unpredictable functionalities, or because the outcome of a computation may be governed by race conditions.
- Last but not least, constraints on *cost* cannot be ignored, be it in terms of money or any other limited resource, such as memory space or available CPU time.

Traditional mainframe systems were proprietary and (essentially) localized; therefore, impact of delays, unforeseen failures, etc. could be considered under the control of the system manager. It was therefore natural, in verification and control of systems, to focus on *functional* behavior entirely.

With the increase in size of computing system and the growing degree of compositionality and distribution, quantitative factors enter the stage:

- calling remote services and transmitting data over the web creates *delays*;
- remote or non-proprietary components are not "deterministic", in the sense that their behavior is uncertain.

Time and *probability* are thus parameters that management of distributed systems must be able to handle; along with both, the *cost* of operations is often subject to restrictions, or its minimization is at least desired. The mathematical treatment of these features in distributed systems is an important challenge, which *MExICo* is addressing; the following describes our activities concerning probabilistic and timed systems. Note that cost optimization is not a current activity but enters the picture in several intended activities.

3.3.2. Probabilistic distributed Systems

Participants: Stefan Haar, Serge Haddad, Claudine Picaronny.

3.3.2.1. Non-sequential probabilistic processes

Practical fault diagnosis requires to select explanations of *maximal likelihood*. For partial-order based diagnosis, this leads therefore to the question what the probability of a given partially ordered execution is. In Benveniste et al. [51], [44], we presented a model of stochastic processes, whose trajectories are partially ordered, based on local branching in Petri net unfoldings; an alternative and complementary model based on Markov fields is developed in [69], which takes a different view on the semantics and overcomes the first model's restrictions on applicability.

Both approaches abstract away from real time progress and randomize choices in *logical* time. On the other hand, the relative speed - and thus, indirectly, the real-time behavior of the system's local processes - are crucial factors determining the outcome of probabilistic choices, even if non-determinism is absent from the system.

In another line of research [55] we have studied the likelihood of occurrence of non-sequential runs under random durations in a stochastic Petri net setting. It remains to better understand the properties of the probability measures thus obtained, to relate them with the models in logical time, and exploit them e.g. in *diagnosis*.

3.3.2.2. Distributed Markov Decision Processes

Participant: Serge Haddad.

Distributed systems featuring non-deterministic and probabilistic aspects are usually hard to analyze and, more specifically, to optimize. Furthermore, high complexity theoretical lower bounds have been established for models like partially observed Markovian decision processes and distributed partially observed Markovian decision processes. We believe that these negative results are consequences of the choice of the models rather than the intrinsic complexity of problems to be solved. Thus we plan to introduce new models in which the associated optimization problems can be solved in a more efficient way. More precisely, we start by studying connection protocols weighted by costs and we look for online and offline strategies for optimizing the mean cost to achieve the protocol. We have been cooperating on this subject with the SUMO team at Inria Rennes; in the joint work [45]; there, we strive to synthesize for a given MDP a control so as to guarantee a specific stationary behavior, rather than - as is usually done - so as to maximize some reward.

3.3.3. Large scale probabilistic systems

Addressing large-scale probabilistic systems requires to face state explosion, due to both the discrete part and the probabilistic part of the model. In order to deal with such systems, different approaches have been proposed:

- Restricting the synchronization between the components as in queuing networks allows to express the steady-state distribution of the model by an analytical formula called a product-form [50].
- Some methods that tackle with the combinatory explosion for discrete-event systems can be generalized to stochastic systems using an appropriate theory. For instance symmetry based methods have been generalized to stochastic systems with the help of aggregation theory [58].
- At last simulation, which works as soon as a stochastic operational semantic is defined, has been adapted to perform statistical model checking. Roughly speaking, it consists to produce a confidence interval for the probability that a random path fulfills a formula of some temporal logic [83] .

We want to contribute to these three axes: (1) we are looking for product-forms related to systems where synchronization are more involved (like in Petri nets), see [9]; (2) we want to adapt methods for discrete-event systems that require some theoretical developments in the stochastic framework and, (3) we plan to address some important limitations of statistical model checking like the expressiveness of the associated logic and the handling of rare events.

3.3.4. Real time distributed systems

Nowadays, software systems largely depend on complex timing constraints and usually consist of many interacting local components. Among them, railway crossings, traffic control units, mobile phones, computer servers, and many more safety-critical systems are subject to particular quality standards. It is therefore becoming increasingly important to look at networks of timed systems, which allow real-time systems to operate in a distributed manner.

Timed automata are a well-studied formalism to describe reactive systems that come with timing constraints. For modeling distributed real-time systems, networks of timed automata have been considered, where the local clocks of the processes usually evolve at the same rate [74] [56]. It is, however, not always adequate to assume that distributed components of a system obey a global time. Actually, there is generally no reason to assume that different timed systems in the networks refer to the same time or evolve at the same rate. Any component is rather determined by local influences such as temperature and workload.

3.3.4.1. Implementation of Real-Time Concurrent Systems

Participants: Thomas Chatain, Stefan Haar, Serge Haddad.

This was one of the tasks of the ANR ImpRo.

Formal models for real-time systems, like timed automata and time Petri nets, have been extensively studied and have proved their interest for the verification of real-time systems. On the other hand, the question of using these models as specifications for designing real-time systems raises some difficulties. One of those comes from the fact that the real-time constraints introduce some artifacts and because of them some syntactically correct models have a formal semantics that is clearly unrealistic. One famous situation is the case of Zeno executions, where the formal semantics allows the system to do infinitely many actions in finite time. But there are other problems, and some of them are related to the distributed nature of the system. These are the ones we address here.

One approach to implementability problems is to formalize either syntactical or behavioral requirements about what should be considered as a reasonable model, and reject other models. Another approach is to adapt the formal semantics such that only realistic behaviors are considered.

These techniques are preliminaries for dealing with the problem of implementability of models. Indeed implementing a model may be possible at the cost of some transformation, which make it suitable for the target device. By the way these transformations may be of interest for the designer who can now use high-level features in a model of a system or protocol, and rely on the transformation to make it implementable.

We aim at formalizing and automating translations that preserve both the timed semantics and the concurrent semantics. This effort is crucial for extending concurrency-oriented methods for logical time, in particular for exploiting partial order properties. In fact, validation and management - in a broad sense - of distributed systems is not realistic *in general* without understanding and control of their real-time dependent features; the link between real-time and logical-time behaviors is thus crucial for many aspects of *MExICo*'s work.

3.3.5. Weighted Automata and Weighted Logics

Participants: Benedikt Bollig, Paul Gastin.

Time and probability are only two facets of quantitative phenomena. A generic concept of adding weights to qualitative systems is provided by the theory of weighted automata [43]. They allow one to treat probabilistic or also reward models in a unified framework. Unlike finite automata, which are based on the Boolean semiring, weighted automata build on more general structures such as the natural or real numbers (equipped with the usual addition and multiplication) or the probabilistic semiring. Hence, a weighted automaton associates with any possible behavior a weight beyond the usual Boolean classification of "acceptance" or "non-acceptance". Automata with weights have produced a well-established theory and come, e.g., with a characterization in terms of rational expressions, which generalizes the famous theorem of Kleene in the unweighted setting. Equipped with a solid theoretical basis, weighted automata finally found their way into numerous application areas such as natural language processing and speech recognition, or digital image compression.

What is still missing in the theory of weighted automata are satisfactory connections with verification-related issues such as (temporal) logic and bisimulation that could lead to a general approach to corresponding satisfiability and model-checking problems. A first step towards a more satisfactory theory of weighted systems was done in [54]. That paper, however, does not give definite answers to all the aforementioned problems. It identifies directions for future research that we will be tackling.

MUTANT Project-Team

3. Research Program

3.1. Real-time Machine Listening

When human listeners are confronted with musical sounds, they rapidly and automatically find their way in the music. Even musically untrained listeners have an exceptional ability to make rapid judgments about music from short examples, such as determining music style, performer, beating, and specific events such as instruments or pitches. Making computer systems capable of similar capabilities requires advances in both music cognition, and analysis and retrieval systems employing signal processing and machine learning.

In a panel session at the 13th National Conference on Artificial Intelligence in 1996, Rodney Brooks (noted figure in robotics) remarked that while automatic speech recognition was a highly researched domain, there had been few works trying to build machines able to understand "non-speech sound". He went further to name this as one of the biggest challenges faced by Artificial Intelligence [41]. More than 15 years have passed. Systems now exist that are able to analyze the contents of music and audio signals and communities such as International Symposium on Music Information Retrieval (MIR) and Sound and Music Computing (SMC) have formed. But we still lack reliable Real-Time machine listening systems.

The first thorough study of machine listening appeared in Eric Scheirer's PhD thesis at MIT Media Lab in 2001 [40] with a focus on low-level listening such as pitch and musical tempo, paving the way for a decade of research. Since the work of Scheirer, the literature has focused on task-dependent methods for machine listening such as pitch estimation, beat detection, structure discovery and more. Unfortunately, the majority of existing approaches are designed for information retrieval on large databases or off-line methods. Whereas the very act of listening is real-time, very little literature exists for supporting real-time machine listening. This argument becomes more clear while looking at the yearly Music Information Retrieval Evaluation eXchange (MIREX), with different retrieval tasks and submitted systems from international institutions, where almost no emphasis exists on real-time machine listening. Most MIR contributions focus on off-line approaches to information retrieval (where the system has access to future data) with less focus on on-line and realtime approaches to information decoding.

On another front, most MIR algorithms suffer from modeling of temporal structures and temporal dynamics specific to music (where most algorithms have roots in speech or biological sequence without correct adoption to temporal streams such as music). Despite tremendous progress using modern signal processing and statistical learning, there is much to be done to achieve the same level of abstract understanding for example in text and image analysis on music data. On another hand, it is important to notice that even untrained listeners are easily able to capture many aspects of formal and symbolic structures from an audio stream in realtime. Realtime machine listening is thus still a major challenge for artificial sciences that should be addressed both on application and theoretical fronts.

In the MuTant project, we focus on realtime and online methods of music information retrieval out of audio signals. One of the primary goals of such systems is to fill in the gap between *signal representation* and *symbolic information* (such as pitch, tempo, expressivity, etc.) contained in music signals. MuTant's current activities focus on two main applications: *score following* or realtime audio-to-score alignment [2], and realtime transcription of music signals [29] with impacts both on signal processing using machine learning techniques and their application in real-world scenarios.

3.2. Synchronous and realtime programming for computer music

The second aspect of an interactive music system is to *react* to extracted high-level and low-level music information based on pre-defined actions. The simplest scenario is *automatic accompaniment*, delegating the interpretation of one or several musical voices to a computer, in interaction with a live solo (or ensemble)

musician(s). The most popular form of such systems is the automatic accompaniment of an orchestral recording with that of a soloist in the classical music repertoire (concertos for example). In the larger context of interactive music systems, the "notes" or musical elements in the accompaniment are replaced by "programs" that are written during the phase of composition and are evaluated in realtime in reaction and relative to musicians' performance. The programs in question here can range from sound playback, to realtime sound synthesis by simulating physical models, and realtime transformation of musician's audio and gesture.

Such musical practice is commonly referred to as the *realtime school* in computer music, developed naturally with the invention of the first score following systems, and led to the invention of the first prototype of realtime digital signal processors [30] and subsequents [34], and the realtime graphical programming environment *Max* for their control [37] at Ircam. With the advent and availability of DSPs in personal computers, integrated realtime event and signal processing graphical language *MaxMSP* was developed [38] at Ircam, which today is the worldwide standard platform for realtime interactive arts programming. This approach to music making was first formalized by composers such as Philippe Manoury and Pierre Boulez, in collaboration with researchers at Ircam, and soon became a standard in musical composition with computers.

Besides realtime performance and implementation issues, little work has underlined the formal aspects of such practices in realtime music programming, in accordance to the long and quite rich tradition of musical notations. Recent progress has convinced both the researcher and artistic bodies that this programming paradigm is close to *synchronous reactive programming languages*, with concrete analogies between both: parallel synchrony and concurrency is equivalent to musical polyphony, periodic sampling to rhythmic patterns, hierarchical structures to micro-polyphonies, and demands for novel hybrid models of time among others. *Antescofo* is therefore an early response to such demands that needs further explorations and studies.

Within the MuTant project, we propose to tackle this aspect of the research within two consecutive lines:

- **Development of a Timed and Synchronous DSL for Real Time Musician-Computer Interaction:** The design of relevant time models and dedicated temporal interactions mechanisms are integrated in the ongoing and continuous development of the *Antescofo* language. The new tools are validated in the production of new musical pieces and other musical applications. This work is performed in strong coupling with composers and performers. The PhD works of José Echeveste (computer science) and Julia Blondeau (composer) take place in this context.
- **Formal Methods:** Failure during an artistic performance should be avoided. This naturally leads to the use of formal methods, like static analysis, verification or test generation, to ensure formally that *Antescofo* programs will behave as expected on stage. The checked properties may also provide some assistance to the composer especially in the context of "non deterministic score" in an interactive framework. The PhD of Clément Poncelet is devoted to these problems.

3.3. Off-the-shelf Operating Systems for Real-time Audio

While operating systems shield the computer hardware from all other software, it provides a comfortable environment for program execution and evades offensive use of hardware by providing various services related to essential tasks. However, integrating discrete and continuous multimedia data demands additional services, especially for real-time processing of continuous-media such as audio and video. To this end interactive systems are sometimes referred to as off-the-shelf operating systems for real-time audio. The difficulty in providing correct real-time services has much to do with human perception. Correctness for real-time audio is more stringent than video because human ear is more sensitive to audio gaps and glitches than human eye is to video jitter [43]. Here we expose the foundations of existing sound and music operating systems and focus on their major drawbacks with regards to today practices.

An important aspect of any real-time operating system is fault-tolerance with regards to short-time failure of continuous-media computation, delivery delay or missing deadlines. Existing multimedia operating systems are soft real-time where missing a deadline does not necessarily lead to system failure and have their roots in pioneering work in [42]. Soft real-time is acceptable in simple applications such as video-on-demand delivery, where initial delay in delivery will not directly lead to critical consequences and can be compensated (general

scheme used for audio-video synchronization), but with considerable consequences for Interactive Systems: Timing failure in interactive systems will heavily affect inter-operability of models of computation, where incorrect ordering can lead to unpredictable and unreliable results. Moreover, interaction between computing and listening machines (both dynamic with respect of internal computation and physical environment) requires tighter and explicit temporal semantics since interaction between physical environment and the system can be continuous and not demand-driven.

Fulfilling timing requirements of continuous media demands explicit use of scheduling techniques. As shown earlier, existing Interactive Music Systems rely on combined event/signal processing. In real-time, scheduling techniques aim at gluing the two engines together with the aim of timely delivery of computations between agents and components, from the physical environment, as well as to hardware components. The first remark in studying existing system is that they all employ static scheduling, whereas interactive computing demands more and more time-aware and context-aware dynamic methods. The scheduling mechanisms are neither aware of time, nor the nature and semantics of computations at stake. Computational elements are considered in a functional manner and reaction and execution requirements are simply ignored. For example, *Max* scheduling mechanisms can delay message delivery when many time-critical tasks are requested within one cycle [38]. *SuperCollider* uses Earliest-Deadline-First (EDF) algorithms and cycles can be simply missed [36]. This situation leads to non-deterministic behavior with deterministic components and poses great difficulties for preservation of underlying techniques, art pieces, and algorithms. The situation has become worse with the demand for nomad physical computing where individual programs and modules are available but no action coordination or orchestration is proposed to design integrated systems. System designers are penalized for expressivity, predictability and reliability of their design despite potentially reliable components.

Existing systems have been successful in programing and executing small system comprised of few programs. However, severe problems arise when scaling from program to system-level for moderate or complex programs leading to unpredictable behavior. Computational elements are considered as functions and reaction and execution requirements are simply ignored. System designers have uniformly chosen to hide timing properties from higher abstractions, and despite its utmost importance in multimedia computing, timing becomes an accident of implementation. This confusing situation for both artists and system designers, is quite similar to the one described in Edward Lee's seminal paper "Computing needs time" stating: "general-purpose computers are increasingly asked to interact with physical processes through integrated media such as audio. [...] and they don't always do it well. The technological basis that engineers have chosen for general-purpose computing [...] does not support these applications well. Changes that ensure this support could improve them and enable many others" [33].

Despite all shortcomings, one of the main advantages of environments such as *Max* and *PureData* to other available systems, and probably the key to their success, is their ability to handle both synchronous processes (such as audio or video delivery and processing) within an asynchronous environment (user and environmental interactions). Besides this fact, multimedia service scheduling at large has a tendency to go more and more towards computing besides mere on-time delivery. This brings in the important question of hybrid scheduling of heterogeneous time and computing models in such environments, a subject that has had very few studies in multimedia processing but studied in areas such simulation applications. We hope to address this issue scientifically by first an explicit study of current challenges in the domain, and second by proposing appropriate methods for such systems. This research is inscribed in the three year ANR project INEDIT coordinated by the team leader (started in September 2012).

PAREO Project-Team

3. Research Program

3.1. Introduction

It is a common claim that rewriting is ubiquitous in computer science and mathematical logic. And indeed the rewriting concept appears from very theoretical settings to very practical implementations. Some extreme examples are the mail system under Unix that uses rules in order to rewrite mail addresses in canonical forms and the transition rules describing the behaviors of tree automata. Rewriting is used in semantics in order to describe the meaning of programming languages [22] as well as in program transformations like, for example, re-engineering of Cobol programs [31]. It is used in order to compute, implicitly or explicitly as in Mathematica or MuPAD, but also to perform deduction when describing by inference rules a logic [18], a theorem prover [20] or a constraint solver [21]. It is of course central in systems making the notion of rule an explicit and first class object, like expert systems, programming languages based on equational logic, algebraic specifications, functional programming and transition systems.

In this context, the study of the theoretical foundations of rewriting have to be continued and effective rewrite based tools should be developed. The extensions of first-order rewriting with higher-order and higher-dimension features are hot topics and these research directions naturally encompass the study of the rewriting calculus, of polygraphs and of their interaction. The usefulness of these concepts becomes more clear when they are implemented and a considerable effort is thus put nowadays in the development of expressive and efficient rewrite based programming languages.

3.2. Rule-based Programming Languages

Programming languages are formalisms used to describe programs, applications, or software which aim to be executed on a given hardware. In principle, any Turing complete language is sufficient to describe the computations we want to perform. However, in practice the choice of the programming language is important because it helps to be effective and to improve the quality of the software. For instance, a web application is rarely developed using a Turing machine or assembly language. By choosing an adequate formalism, it becomes easier to reason about the program, to analyze, certify, transform, optimize, or compile it. The choice of the programming language also has an impact on the quality of the software. By providing high-level constructs as well as static verifications, like typing, we can have an impact on the software design, allowing more expressiveness, more modularity, and a better reuse of code. This also improves the productivity of the programmer, and contributes to reducing the presence of errors.

The quality of a programming language depends on two main factors. First, the *intrinsic design*, which describes the programming model, the data model, the features provided by the language, as well as the semantics of the constructs. The second factor is the programmer and the application which is targeted. A language is not necessarily good for a given application if the concepts of the application domain cannot be easily manipulated. Similarly, it may not be good for a given person if the constructs provided by the language are not correctly understood by the programmer.

In the *Pareo* group we target a population of programmers interested in improving the long-term maintainability and the quality of their software, as well as their efficiency in implementing complex algorithms. Our privileged domain of application is large since it concerns the development of *transformations*. This ranges from the transformation of textual or structured documents such as XML, to the analysis and the transformation of programs and models. This also includes the development of tools such as theorem provers, proof assistants, or model checkers, where the transformations of proofs and the transitions between states play a crucial role. In that context, the *expressiveness* of the programming language is important. Indeed, complex encodings into low level data structures should be avoided, in contrast to high level notions such as abstract types and transformation rules that should be provided.

It is now well established that the notions of *term* and *rewrite rule* are two universal abstractions well suited to model tree based data types and the transformations that can be done upon them. Over the last ten years we have developed a strong experience in designing and programming with rule based languages [23], [14], [12]. We have introduced and studied the notion of *strategy* [13], which is a way to control how the rules should be applied. This provides the separation which is essential to isolate the logic and to make the rules reusable in different contexts.

To improve the quality of programs, it is also essential to have a clear description of their intended behaviors. For that, the *semantics* of the programming language should be formally specified.

There is still a lot of progress to be done in these directions. In particular, rule based programming can be made even more expressive by extending the existing matching algorithms to context-matching or to new data structures such as graphs or polygraphs. New algorithms and implementation techniques have to be found to improve the efficiency and make the rule based programming approach effective on large problems. Separating the rules from the control is very important. This is done by introducing a language for describing strategies. We still have to invent new formalisms and new strategy primitives which are both expressive enough and theoretically well grounded. A challenge is to find a good strategy language we can reason about, to prove termination properties for instance.

On the static analysis side, new formalized typing algorithms are needed to properly integrate rule based programming into already existing host languages such as Java. The notion of traversal strategy merits to be better studied in order to become more flexible and still provide a guarantee that the result of a transformation is correctly typed.

3.3. Rewriting Calculus

The huge diversity of the rewriting concept is obvious and when one wants to focus on the underlying notions, it becomes quickly clear that several technical points should be settled. For example, what kind of objects are rewritten? Terms, graphs, strings, sets, multisets, others? Once we have established this, what is a rewrite rule? What is a left-hand side, a right-hand side, a condition, a context? And then, what is the effect of a rule application? This leads immediately to defining more technical concepts like variables in bound or free situations, substitutions and substitution application, matching, replacement; all notions being specific to the kind of objects that have to be rewritten. Once this is solved one has to understand the meaning of the application of a set of rules on (classes of) objects. And last but not least, depending on the intended use of rewriting, one would like to define an induced relation, or a logic, or a calculus.

In this very general picture, we have introduced a calculus whose main design concept is to make all the basic ingredients of rewriting explicit objects, in particular the notions of rule *application* and *result*. We concentrate on *term* rewriting, we introduce a very general notion of rewrite rule and we make the rule application and result explicit concepts. These are the basic ingredients of the *rewriting*- or ρ-calculus whose originality comes from the fact that terms, rules, rule application and application strategies are all treated at the object level (a rule can be applied on a rule for instance).

The λ-calculus is usually put forward as the abstract computational model underlying functional programming. However, modern functional programming languages have pattern-matching features which cannot be directly expressed in the λ-calculus. To palliate this problem, pattern-calculi [28], [25], [19] have been introduced. The rewriting calculus is also a pattern calculus that combines the expressiveness of pure functional calculi and algebraic term rewriting. This calculus is designed and used for logical and semantical purposes. It could be equipped with powerful type systems and used for expressing the semantics of rule based as well as object oriented languages. It allows one to naturally express exception handling mechanisms and elaborated rewriting strategies. It can be also extended with imperative features and cyclic data structures.

The study of the rewriting calculus turns out to be extremely successful in terms of fundamental results and of applications [16]. Different instances of this calculus together with their corresponding type systems have been proposed and studied. The expressive power of this calculus was illustrated by comparing it with similar

formalisms and in particular by giving a typed encoding of standard strategies used in first-order rewriting and classical rewrite based languages like *ELAN* and *Tom*.

PARKAS Project-Team

3. Research Program

3.1. Presentation and originality of the PARKAS team

Our project is founded on our expertise in three complementary domains: (1) synchronous functional programming and its extensions to deal with features such as communication with bounded buffers and dynamic process creation; (2) mathematical models for synchronous circuits; (3) compilation techniques for synchronous languages and optimizing/parallelizing compilers.

A strong point of the team is its experience and investment in the development of languages and compilers. Members of the team also have direct collaborations for several years with major industrial companies in the field and several of our results are integrated in successful products. Our main results are briefly summarized below.

3.1.1. Synchronous functional programming

In [30], Paul Caspi and Marc Pouzet introduced *synchronous Kahn networks* as those Kahn networks that can be statically scheduled and executed with bounded buffers. This was the origin of the language LUCID SYNCHRONE, [00] an ML extension of the synchronous language LUSTRE with higher-order features, dedicated type systems (clock calculus as a type system [30], [41], initialization analysis [42] and causality analysis [44]). The language integrates original features that are not found in other synchronous languages: such as combinations of data flow, control flow, hierarchical automata and signals [40], [39], and modular code generation [31], [26].

In 2000, Marc Pouzet started to collaborate with the SCADE team of Esterel-Technologies on the design of a new version of SCADE. [0] Several features of LUCID SYNCHRONE are now integrated into SCADE 6, which has been distributed since 2008, including the programming constructs `merge`, `reset`, the clock calculus and the type system. Several results have been developed jointly with Jean-Louis Colaço and Bruno Pagano from Esterel-Technologies, such as ways of combining data-flow and hierarchical automata, and techniques for their compilation, initialization analysis, etc.

Dassault-Systèmes (Grenoble R&D center, part of Delmia-automation) developed the language LCM, a variant of LUCID SYNCHRONE that is used for the simulation of factories. LCM follows closely the principles and programming constructs of LUCID SYNCHRONE (higher-order, type inference, mix of data-flow and hierarchical automata). The team in Grenoble is integrating this development into a new compiler for the language Modelica. [0]

In parallel, the goal of REACTIVEML[0] was to integrate a synchronous concurrency model into an existing ML language, with no restrictions on expressiveness, so as to program a large class of reactive systems, including efficient simulations of millions of communicating processes (e.g., sensor networks), video games with many interactions, physical simulations, etc. For such applications, the synchronous model simplifies system design and implementation, but the expressiveness of the algorithmic part of the language is just as essential, as is the ability to create or stop a process dynamically.

The development of REACTIVEML was started by Louis Mandel during his PhD thesis [55], [53] and is ongoing. The language extends OCAML[0] with Esterel-like synchronous primitives — synchronous composition, broadcast communication, pre-emption/suspension — applying the solution of Boussinot [27] to solve causality issues.

[0] http://www.di.ens.fr/~pouzet/lucid-synchrone
[0] The name is a reference to Lustre which stands for "Lucid Synchrone et Temps réel".
[0] http://www.esterel-technologies.com/products/scade-suite/
[0] http://www.3ds.com/products/catia/portfolio/dymola/overview/
[0] http://rml.lri.fr/
[0] More precisely a subset of OCAML without objects or functors.

Several open problems have been solved by Louis Mandel: the interaction between ML features (higher-order) and reactive constructs with a proper type system; efficient simulation that avoids busy waiting. The latter problem is particularly difficult in synchronous languages because of possible reactions to the absence of a signal. In the REACTIVEML implementation, there is no busy waiting: inactive processes have no impact on the overall performance. It turns out that this enables REACTIVEML to simulate millions of (logical) parallel processes and to compete with the best event-driven simulators [56].

REACTIVEML has been used for simulating routing protocols in ad-hoc networks [52] and large scale sensor networks [67]. The designer benefits from a real programming language that gives precise control of the level of simulation (e.g., each network layer up to the MAC layer) and programs can be connected to models of the physical environment programmed with LUTIN [66]. REACTIVEML is used since 2006 by the synchronous team at VERIMAG, Grenoble (in collaboration with France-Telecom) for the development of low-consumption routing protocols in sensor networks.

3.1.2. Relaxing synchrony with buffer communication

In the data-flow synchronous model, the clock calculus is a static analysis that ensures execution in bounded memory. It checks that the values produced by a node are instantaneously consumed by connected nodes (synchronous constraint). To program Kahn process networks with bounded buffers (as in video applications), it is thus necessary to explicitly place nodes that implement buffers. The buffers sizes and the clocks at which data must be read or written have to be computed manually. In practice, it is done with simulation or successive tries and errors. This task is difficult and error prone. The aim of the n-synchronous model is to automatically compute at compile time these values while insuring the absence of deadlock.

Technically, it allows processes to be composed whenever they can be synchronized through a bounded buffer [32], [33]. The new flexibility is obtained by relaxing the clock calculus by replacing the equality of clocks by a sub-typing rule. The result is a more expressive language which still offers the same guarantees as the original. The first version of the model was based on clocks represented as ultimately periodic binary words [73]. It was algorithmically expensive and limited to periodic systems. In [37], an abstraction mechanism is proposed which permits direct reasoning on sets of clocks that are defined as a rational slope and two shifts. An implementation of the n-synchronous model, named LUCY-N, was developed in 2009 [54], as was a formalization of the theory in COQ [38]. We also worked on low-level compiler and runtime support to parallelize the execution of relaxed synchronous systems, proposing a portable intermediate language and runtime library called ERBIUM [57].

This work started as a collaboration between Marc Pouzet (LIP6, Paris, then LRI and Inria Proval, Orsay), Marc Duranton (Philips Research then NXP, Eindhoven), Albert Cohen (Inria Alchemy, Orsay) and Christine Eisenbeis (Inria Alchemy, Orsay) on the real-time programming of video stream applications in set-top boxes. It was significantly extended by Louis Mandel and Florence Plateau during her PhD thesis [61] (supervised by Marc Pouzet and Louis Mandel). Low-level support has been investigated with Cupertino Miranda, Philippe Dumont (Inria Alchemy, Orsay) and Antoniu Pop (Mines ParisTech). Further directions of research and experimentation have been and are being followed through the theses of Léonard Gérard, Adrien Guatto and Nhat Minh Lê.

3.1.3. Polyhedral compilation and optimizing compilers

Despite decades of progress, the best parallelizing and optimizing compilers still fail to extract parallelism and to perform the necessary optimizations to harness multi-core processors and their complex memory hierarchies. *Polyhedral compilation* aims at facilitating the construction of more effective optimization and parallelization algorithms. It captures the flow of data between individual instances of statements in a loop nest, allowing to accurately model the behavior of the program and represent complex parallelizing and optimizing transformations. Affine multidimensional scheduling is one of the main tools in polyhedral compilation [45]. Albert Cohen, in collaboration with Cédric Bastoul, Sylvain Girbal, Nicolas Vasilache, Louis-Noël Pouchet and Konrad Trifunovic (LRI and Inria Alchemy, Orsay) has contributed to a large number of research, development and transfer activities in this area.

The relation between polyhedral compilation and data-flow synchrony has been identified through data-flow array languages [51], [50], [68], [46] and the study of the scheduling and mapping algorithms for these languages. We would like to deepen the exploration of this link, embedding polyhedral techniques into the compilation flow of data-flow, relaxed synchronous languages.

Our previous work led to the design of a theoretical and algorithmic framework rooted in the polyhedral model of compilation, and to the implementation of a set of tools based on production compilers (Open64, GCC) and source-to-source prototypes (PoCC, http://pocc.sourceforge.net). We have shown that not only does this framework simplify the problem of building complex loop nest optimizations, but also that it scales to real-world benchmarks [34], [47], [64], [63]. The polyhedral model has finally evolved into a mature, production-ready approach to solve the challenges of maximizing the scalability and efficiency of loop-based computations on a variety of high performance and embedded targets.

After an initial experiment with Open64 [35], [34], we ported these techniques to GCC [62], [70], [69] and LLVM [49], applying them to multi-level parallelization and optimization problems, including vectorization and exploitation of thread-level parallelism. Independently, we made significant progress in the design of effective optimization heuristics, working on the interactions between the semantics of the compiler's intermediate representation and the structure of the optimization space [64], [63], [65], [23], [60]. These results open opportunities for complex optimizations that target larger problems, such as the scheduling and placement of process networks, or the offloading of computational kernels to hardware accelerators (such as GPUs). A new framework has been designed, centered on the Integer Set Library (isl, http://freecode.com/projects/isl) and implemented through multiple compiler interfaces (Graphite in GCC, Polly in LLVM) and a source-to-source research compiler (PPCG) [72], [36], [48], [71]. This new framework underlies our collaborative research activities in the CARP and COPCAMS European projects, as well as emerging transfer projects through the TETRACOM European coordination action and bilateral industry contracts in preparation.

3.1.4. Automatic compilation of high performance circuits

For both cost and performance reasons, computing systems tightly couple parts realized in hardware with parts realized in software. The boundary between hardware and software keeps moving with the underlying technology and the external economic pressure. Moreover, thanks to FPGA technology, hardware itself has become programmable. There is now a pressing need from industry for hardware/software co-design, and for tools which automatically turn software code into hardware circuits, or more usually, into hybrid code that simultaneously targets GPUs, multiple cores, encryption ASICs, and other specialized chips.

Departing from customary C-to-VHDL compilation, we trust that sharper results can be achieved from source programs that specify bit-wise time/space behavior in a rigorous synchronous language, rather than just the I/O behavior in some (ill-specified) subset of C. This specification allows the designer to also program the (asynchronous) environment in which to operate the entire system, and to profile/measure/control each variable of the design.

At any time, the designer can edit a single specification of the system, from which both the software and the hardware are automatically compiled, and guaranteed to be compatible. Once correct (functionally and with respect to the behavioral specification), the application can be automatically deployed (and tested) on a hard/soft hybrid co-design support.

Key aspects of the advocated methodology were validated by Jean Vuillemin in the design of a PAL2HDTV video sampler [58], [59]. The circuit was automatically compiled from a synchronous source specification, decorated and guided by a few key hints to the hardware back-end, that targetted an FPGA running at real-time video specifications: a tightly-packed highly-efficient design at 240MHz, generated 100% automatically from the application specification source code, and including all run-time/debug/test/validate ancillary software. It was subsequently commercialized on FPGA by LetItWave, and then on ASIC by Zoran. This successful experience underlines our research perspectives on parallel synchronous programming.

PARSIFAL Project-Team

3. Research Program

3.1. General overview

There are two broad approaches for computational specifications. In the *computation as model* approach, computations are encoded as mathematical structures containing nodes, transitions, and state. Logic is used to *describe* these structures, that is, the computations are used as models for logical expressions. Intensional operators, such as the modals of temporal and dynamic logics or the triples of Hoare logic, are often employed to express propositions about the change in state.

The *computation as deduction* approach, in contrast, expresses computations logically, using formulas, terms, types, and proofs as computational elements. Unlike the model approach, general logical apparatus such as cut-elimination or automated deduction becomes directly applicable as tools for defining, analyzing, and animating computations. Indeed, we can identify two main aspects of logical specifications that have been very fruitful:

- *Proof normalization*, which treats the state of a computation as a proof term and computation as normalization of the proof terms. General reduction principles such as β-reduction or cut-elimination are merely particular forms of proof normalization. Functional programming is based on normalization [64], and normalization in different logics can justify the design of new and different functional programming languages [38].
- *Proof search*, which views the state of a computation as a a structured collection of formulas, known as a *sequent*, and proof search in a suitable sequent calculus as encoding the dynamics of the computation. Logic programming is based on proof search [70], and different proof search strategies can be used to justify the design of new and different logic programming languages [68].

While the distinction between these two aspects is somewhat informal, it helps to identify and classify different concerns that arise in computational semantics. For instance, confluence and termination of reductions are crucial considerations for normalization, while unification and strategies are important for search. A key challenge of computational logic is to find means of uniting or reorganizing these apparently disjoint concerns.

An important organizational principle is structural proof theory, that is, the study of proofs as syntactic, algebraic and combinatorial objects. Formal proofs often have equivalences in their syntactic representations, leading to an important research question about *canonicity* in proofs – when are two proofs "essentially the same?" The syntactic equivalences can be used to derive normal forms for proofs that illuminate not only the proofs of a given formula, but also its entire proof search space. The celebrated *focusing* theorem of Andreoli [39] identifies one such normal form for derivations in the sequent calculus that has many important consequences both for search and for computation. The combinatorial structure of proofs can be further explored with the use of *deep inference*; in particular, deep inference allows access to simple and manifestly correct cut-elimination procedures with precise complexity bounds.

Type theory is another important organizational principle, but most popular type systems are generally designed for either search or for normalization. To give some examples, the Coq system [76] that implements the Calculus of Inductive Constructions (CIC) is designed to facilitate the expression of computational features of proofs directly as executable functional programs, but general proof search techniques for Coq are rather primitive. In contrast, the Twelf system [72] that is based on the LF type theory (a subsystem of the CIC), is based on relational specifications in canonical form (*i.e.*, without redexes) for which there are sophisticated automated reasoning systems such as meta-theoretic analysis tools, logic programming engines, and inductive theorem provers. In recent years, there has been a push towards combining search and normalization in the same type-theoretic framework. The Beluga system [73], for example, is an extension of the LF type theory with a purely computational meta-framework where operations on inductively defined LF objects can be expressed as functional programs.

The Parsifal team investigates both the search and the normalization aspects of computational specifications using the concepts, results, and insights from proof theory and type theory.

3.2. Inductive and co-inductive reasoning

The team has spent a number of years in designing a strong new logic that can be used to reason (inductively and co-inductively) on syntactic expressions containing bindings. This work is based on earlier work by McDowell, Miller, and Tiu [66] [65] [71] [77], and on more recent work by Gacek, Miller, and Nadathur [3] [52]. The Parsifal team, along with our colleagues in Minneapolis, Canberra, Singapore, and Cachen, have been building two tools that exploit the novel features of this logic. These two systems are the following.

- Abella, which is an interactive theorem prover for the full logic.
- Bedwyr, which is a model checker for the "finite" part of the logic.

We have used these systems to provide formalize reasoning of a number of complex formal systems, ranging from programming languages to the λ-calculus and π-calculus.

During 2014, the Abella system has been extended with a number of new features. A number of new significant examples have been implemented in Abella and an extensive tutorial for it has been written [31].

3.3. Developing a foundational approach to defining proof evidence

The team is developing a framework for defining the semantics of proof evidence. With this framework, implementers of theorem provers can output proof evidence in a format of their choice: they will only need to be able to formally define that evidence's semantics. With such semantics provided, proof checkers can then check alleged proofs for correctness. Thus, anyone who needs to trust proofs from various provers can put their energies into designing trustworthy checkers that can execute the semantic specification.

In order to provide our framework with the flexibility that this ambitious plan requires, we have based our design on the most recent advances within the theory of proofs. For a number of years, various team members have been contributing to the design and theory of *focused proof systems* [40] [42] [44] [45] [55] [62] [63] and we have adopted such proof systems as the corner stone for our framework.

We have also been working for a number of years on the implementation of computational logic systems, involving, for example, both unification and backtracking search. As a result, we are also building an early and reference implementation of our semantic definitions.

3.4. Deep inference

Deep inference [57], [59] is a novel methodology for presenting deductive systems. Unlike traditional formalisms like the sequent calculus, it allows rewriting of formulas deep inside arbitrary contexts. The new freedom for designing inference rules creates a richer proof theory. For example, for systems using deep inference, we have a greater variety of normal forms for proofs than in sequent calculus or natural deduction systems. Another advantage of deep inference systems is the close relationship to categorical proof theory. Due to the deep inference design one can directly read off the morphism from the derivations. There is no need for a counter-intuitive translation.

The following research problems are investigated by members of the Parsifal team:

- Find deep inference system for richer logics. This is necessary for making the proof theoretic results of deep inference accessible to applications as they are described in the previous sections of this report.
- Investigate the possibility of focusing proofs in deep inference. As described before, focusing is a way to reduce the non-determinism in proof search. However, it is well investigated only for the sequent calculus. In order to apply deep inference in proof search, we need to develop a theory of focusing for deep inference.

3.5. Proof nets and atomic flows

Proof nets and atomic flows are abstract (graph-like) presentations of proofs such that all "trivial rule permutations" are quotiented away. Ideally the notion of proof net should be independent from any syntactic formalism, but most notions of proof nets proposed in the past were formulated in terms of their relation to the sequent calculus. Consequently we could observe features like “boxes” and explicit “contraction links”. The latter appeared not only in Girard’s proof nets [54] for linear logic but also in Robinson’s proof nets [74] for classical logic. In this kind of proof nets every link in the net corresponds to a rule application in the sequent calculus.

Only recently, due to the rise of deep inference, new kinds of proof nets have been introduced that take the formula trees of the conclusions and add additional “flow-graph” information (see e.g., [5], [4] and [58]. On one side, this gives new insights in the essence of proofs and their normalization. But on the other side, all the known correctness criteria are no longer available.

This directly leads to the following research questions investigated by members of the Parsifal team:

- Finding (for classical logic) a notion of proof nets that is deductive, i.e., can effectively be used for doing proof search. An important property of deductive proof nets must be that the correctness can be checked in linear time. For the classical logic proof nets by Lamarche and Straßburger [5] this takes exponential time (in the size of the net).
- Studying the normalization of proofs in classical logic using atomic flows. Although there is no correctness criterion they allow to simplify the normalization procedure for proofs in deep inference, and additionally allow to get new insights in the complexity of the normalization.

PI.R2 Project-Team

3. Research Program

3.1. Proof theory and the Curry-Howard correspondence

3.1.1. Proofs as programs

Proof theory is the branch of logic devoted to the study of the structure of proofs. An essential contributor to this field is Gentzen [51] who developed in 1935 two logical formalisms that are now central to the study of proofs. These are the so-called "natural deduction", a syntax that is particularly well-suited to simulate the intuitive notion of reasoning, and the so-called "sequent calculus", a syntax with deep geometric properties that is particularly well-suited for proof automation.

Proof theory gained a remarkable importance in computer science when it became clear, after genuine observations first by Curry in 1958 [44], then by Howard and de Bruijn at the end of the 60's [54], [66], that proofs had the very same structure as programs: for instance, natural deduction proofs can be identified as typed programs of the ideal programming language known as λ-calculus.

This proofs-as-programs correspondence has been the starting point to a large spectrum of researches and results contributing to deeply connect logic and computer science. In particular, it is from this line of work that Coquand's Calculus of Constructions [41] stemmed out – a formalism that is both a logic and a programming language and that is at the source of the Coq system [64].

3.1.2. Towards the calculus of constructions

The λ-calculus, defined by Church [40], is a remarkably succinct model of computation that is defined via only three constructions (abstraction of a program with respect to one of its parameters, reference to such a parameter, application of a program to an argument) and one reduction rule (substitution of the formal parameter of a program by its effective argument). The λ-calculus, which is Turing-complete, i.e. which has the same expressiveness as a Turing machine (there is for instance an encoding of numbers as functions in λ-calculus), comes with two possible semantics referred to as call-by-name and call-by-value evaluations. Of these two semantics, the first one, which is the simplest to characterise, has been deeply studied in the last decades [37].

For explaining the Curry-Howard correspondence, it is important to distinguish between intuitionistic and classical logic: following Brouwer at the beginning of the 20^{th} century, classical logic is a logic that accepts the use of reasoning by contradiction while intuitionistic logic proscribes it. Then, Howard's observation is that the proofs of the intuitionistic natural deduction formalism exactly coincide with programs in the (simply typed) λ-calculus.

A major achievement has been accomplished by Martin-Löf who designed in 1971 a formalism, referred to as modern type theory, that was both a logical system and a (typed) programming language [60].

In 1985, Coquand and Huet [41], [42] in the Formel team of Inria-Rocquencourt explored an alternative approach based on Girard-Reynolds' system F [52], [63]. This formalism, called the Calculus of Constructions, served as logical foundation of the first implementation of Coq in 1984. Coq was called CoC at this time.

3.1.3. The Calculus of Inductive Constructions

The first public release of CoC dates back to 1989. The same project-team developed the programming language Caml (nowadays called OCaml and coordinated by the Gallium team) that provided the expressive and powerful concept of algebraic data types (a paragon of it being the type of list). In CoC, it was possible to simulate algebraic data types, but only through a not-so-natural not-so-convenient encoding.

In 1989, Coquand and Paulin [43] designed an extension of the Calculus of Constructions with a generalisation of algebraic types called inductive types, leading to the Calculus of Inductive Constructions (CIC) that started to serve as a new foundation for the Coq system. This new system, which got its current definitive name Coq, was released in 1991.

In practice, the Calculus of Inductive Constructions derives its strength from being both a logic powerful enough to formalise all common mathematics (as set theory is) and an expressive richly-typed functional programming language (like ML but with a richer type system, no effects and no non-terminating functions).

3.2. The development of Coq

Since 1984, about 40 persons have contributed to the development of Coq, out of which 7 persons have contributed to bring the system to the place it is now. First Thierry Coquand through his foundational theoretical ideas, then Gérard Huet who developed the first prototypes with Thierry Coquand and who headed the Coq group until 1998, then Christine Paulin who was the main actor of the system based on the CIC and who headed the development group from 1998 to 2006. On the programming side, important steps were made by Chet Murthy who raised Coq from the prototypical state to a reasonably scalable system, Jean-Christophe Filliâtre who turned to concrete the concept of a small trustful certification kernel on which an arbitrary large system can be set up, Bruno Barras and Hugo Herbelin who, among other extensions, reorganised Coq on a new smoother and more uniform basis able to support a new round of extensions for the next decade.

The development started from the Formel team at Rocquencourt but, after Christine Paulin got a position in Lyon, it spread to École Normale Supérieure de Lyon. Then, the task force there globally moved to the University of Orsay when Christine Paulin got a new position there. On the Rocquencourt side, the part of Formel involved in ML moved to the Cristal team (now Gallium) and Formel got renamed into Coq. Gérard Huet left the team and Christine Paulin started to head a Coq team bilocalised at Rocquencourt and Orsay. Gilles Dowek became the head of the team which was renamed into LogiCal. Following Gilles Dowek who got a position at École Polytechnique, LogiCal moved to the new Inria Saclay research center. It then split again, giving birth to ProVal. At the same time, the Marelle team (formerly Lemme, formerly Croap) which has been a long partner of the Formel team, invested more and more energy in both the formalisation of mathematics in Coq and in user interfaces for Coq.

After various other spreadings resulting from where the wind pushed former PhD students, the development of Coq got multi-site with the development now realised by employees of Inria, the CNAM and Paris 7.

We next briefly describe the main components of Coq.

3.2.1. The underlying logic and the verification kernel

The architecture adopts the so-called de Bruijn principle: the well-delimited *kernel* of Coq ensures the correctness of the proofs validated by the system. The kernel is rather stable with modifications tied to the evolution of the underlying Calculus of Inductive Constructions formalism. The kernel includes an interpreter of the programs expressible in the CIC and this interpreter exists in two flavours: a customisable lazy evaluation machine written in OCaml and a call-by-value bytecode interpreter written in C dedicated to efficient computations. The kernel also provides a module system.

3.2.2. Programming and specification languages

The concrete user language of Coq, called *Gallina*, is a high-level language built on top of the CIC. It includes a type inference algorithm, definitions by complex pattern-matching, implicit arguments, mathematical notations and various other high-level language features. This high-level language serves both for the development of programs and for the formalisation of mathematical theories. Coq also provides a large set of commands. Gallina and the commands together forms the *Vernacular* language of Coq.

3.2.3. Libraries

Libraries are written in the vernacular language of Coq. There are libraries for various arithmetical structures and various implementations of numbers (Peano numbers, implementation of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ with binary digits, implementation of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ using machine words, axiomatisation of $\mathbb{R}$). There are libraries for lists, list of a specified length, sorts, and for various implementations of finite maps and finite sets. There are libraries on relations, sets, orders.

3.2.4. Tactics

The tactics are the methods available to conduct proofs. This includes the basic inference rules of the CIC, various advanced higher level inference rules and all the automation tactics. Regarding automation, there are tactics for solving systems of equations, for simplifying ring or field expressions, for arbitrary proof search, for semi-decidability of first-order logic and so on. There is also a powerful and popular untyped scripting language for combining tactics into more complex tactics.

Note that all tactics of Coq produce proof certificates that are checked by the kernel of Coq. As a consequence, possible bugs in proof methods do not hinder the confidence in the correctness of the Coq checker. Note also that the CIC being a programming language, tactics can be written (and certified) in the own language of Coq if needed.

3.2.5. Extraction

Extraction is a component of Coq that maps programs (or even computational proofs) of the CIC to functional programs (in OCaml, Scheme or Haskell). Especially, a program certified by Coq can further be extracted to a program of a full-fledged programming language then benefiting of the efficient compilation, linking tools, profiling tools, ... of the target software.

3.3. Dependently typed programming languages

Dependently typed programming (shortly DTP) is an emerging concept referring to the diffuse and broadening tendency to develop programming languages with type systems able to express program properties finer than the usual information of simply belonging to specific data-types. The type systems of dependently-typed programming languages allow to express properties *dependent* of the input and the output of the program (for instance that a sorting program returns a list of same size as its argument). Typical examples of such languages were the Cayenne language, developed in the late 90's at Chalmers University in Sweden and the DML language developed at Boston. Since then, various new tools have been proposed, either as typed programming languages whose types embed equalities (Ωmega at Portland, ATS at Boston, ...) or as hybrid logic/programming frameworks (Agda at Chalmers University, Twelf at Carnegie, Delphin at Yale, OpTT at U. Iowa, Epigram at Nottingham, ...).

DTP contributes to a general movement leading to the fusion between logic and programming. Coq, whose language is both a logic and a programming language which moreover can be extracted to pure ML code plays a role in this movement and some frameworks for DTP have been proposed on top of Coq (Concoqtion at Rice and Colorado, Ynot at Harvard, Why in the ProVal team at Inria). It also connects to Hoare logic, providing frameworks where pre- and post-conditions of programs are tied with the programs.

DTP approached from the programming language side generally benefits of a full-fledged language (e.g. supporting effects) with efficient compilation. DTP approached from the logic side generally benefits of an expressive specification logic and of proof methods so as to certify the specifications. The weakness of the approach from logic however is generally the weak support for effects or partial functions.

3.3.1. Type-checking and proof automation

In between the decidable type systems of conventional data-types based programming languages and the full expressiveness of logically undecidable formulae, an active field of research explores a spectrum of decidable or semi-decidable type systems for possible use in dependently typed programming languages. At the beginning of the spectrum, this includes, for instance, the system F's extension ML_F of the ML type

system or the generalisation of abstract data types with type constraints (G.A.D.T.) such as found in the Haskell programming language. At the other side of the spectrum, one finds arbitrary complex type specification languages (e.g. that a sorting function returns a list of type "sorted list") for which more or less powerful proof automation tools exist – generally first-order ones.

3.4. Around and beyond the Curry-Howard correspondence

For two decades, the Curry-Howard correspondence has been limited to the intuitionistic case but since 1990, an important stimulus spurred on the community following Griffin's discovery that this correspondence was extensible to classical logic. The community then started to investigate unexplored potential connections between computer science and logic. One of these fields is the computational understanding of Gentzen's sequent calculus while another one is the computational content of the axiom of choice.

3.4.1. Control operators and classical logic

Indeed, a significant extension of the Curry-Howard correspondence has been obtained at the beginning of the 90's thanks to the seminal observation by Griffin [53] that some operators known as control operators were typable by the principle of double negation elimination ($\neg\neg A \Rightarrow A$), a principle that enables classical reasoning.

Control operators are used to jump from one location of a program to another. They were first considered in the 60's by Landin [58] and Reynolds [62] and started to be studied in an abstract way in the 80's by Felleisen *et al* [49], leading to Parigot's $\lambda\mu$-calculus [61], a reference calculus that is in close Curry-Howard correspondence with classical natural deduction. In this respect, control operators are fundamental pieces to establish a full connection between proofs and programs.

3.4.2. Sequent calculus

The Curry-Howard interpretation of sequent calculus started to be investigated at the beginning of the 90's. The main technicality of sequent calculus is the presence of *left introduction* inference rules, for which two kinds of interpretations are applicable. The first approach interprets left introduction rules as construction rules for a language of patterns but it does not really address the problem of the interpretation of the implication connective. The second approach, started in 1994, interprets left introduction rules as evaluation context formation rules. This line of work led in 2000 to the design by Hugo Herbelin and Pierre-Louis Curien of a symmetric calculus exhibiting deep dualities between the notion of programs and evaluation contexts and between the standard notions of call-by-name and call-by-value evaluation semantics.

3.4.3. Abstract machines

Abstract machines came as an intermediate evaluation device, between high-level programming languages and the computer microprocessor. The typical reference for call-by-value evaluation of λ-calculus is Landin's SECD machine [57] and Krivine's abstract machine for call-by-name evaluation [56], [55]. A typical abstract machine manipulates a state that consists of a program in some environment of bindings and some evaluation context traditionally encoded into a "stack".

3.4.4. Delimited control

Delimited control extends the expressiveness of control operators with effects: the fundamental result here is a completeness result by Filinski [50]: any side-effect expressible in monadic style (and this covers references, exceptions, states, dynamic bindings, ...) can be simulated in λ-calculus equipped with delimited control.

POLSYS Project-Team

3. Research Program

3.1. Introduction

Polynomial system solving is a fundamental problem in Computer Algebra with many applications in cryptography, robotics, biology, error correcting codes, signal theory, ... Among all available methods for solving polynomial systems, computation of Gröbner bases remains one of the most powerful and versatile method since it can be applied in the continuous case (rational coefficients) as well as in the discrete case (finite fields). Gröbner bases are also a building blocks for higher level algorithms who compute real sample points in the solution set of polynomial systems, decide connectivity queries and quantifier elimination over the reals. The major challenge facing the designer or the user of such algorithms is the intrinsic exponential behaviour of the complexity for computing Gröbner bases. The current proposal is an attempt to tackle these issues in a number of different ways: improve the efficiency of the fundamental algorithms (even when the complexity is exponential), develop high performance implementation exploiting parallel computers, and investigate new classes of structured algebraic problems where the complexity drops to polynomial time.

3.2. Fundamental Algorithms and Structured Systems

Participants: Jean-Charles Faugère, Mohab Safey El Din, Elias Tsigaridas, Guénaël Renault, Dongming Wang, Jérémy Berthomieu, Jules Svartz, Louise Huot, Thibaut Verron.

Efficient algorithms F_4/F_5[0] for computing the Gröbner basis of a polynomial system rely heavily on a connection with linear algebra. Indeed, these algorithms reduce the Gröbner basis computation to a sequence of Gaussian eliminations on several submatrices of the so-called Macaulay matrix in some degree. Thus, we expect to improve the existing algorithms by
(i) developing dedicated linear algebra routines performing the Gaussian elimination steps: this is precisely the objective 2 described below;
(ii) generating smaller or simpler matrices to which we will apply Gaussian elimination.
We describe here our goals for the latter problem. First, we focus on algorithms for computing a Gröbner basis of *general polynomial systems*. Next, we present our goals on the development of dedicated algorithms for computing Gröbner bases of *structured polynomial systems* which arise in various applications.

Algorithms for general systems. Several degrees of freedom are available to the designer of a Gröbner basis algorithm to generate the matrices occurring during the computation. For instance, it would be desirable to obtain matrices which would be almost triangular or very sparse. Such a goal can be achieved by considering various interpretations of the F_5 algorithm with respect to different monomial orderings. To address this problem, the tight complexity results obtained for F_5 will be used to help in the design of such a general algorithm. To illustrate this point, consider the important problem of solving boolean polynomial systems; it might be interesting to preserve the sparsity of the original equations and, at the same time, using the fact that overdetermined systems are much easier to solve.

Algorithms dedicated to *structured* polynomial systems. A complementary approach is to exploit the structure of the input polynomials to design specific algorithms. Very often, problems coming from applications are not random but are highly structured. The specific nature of these systems may vary a lot: some polynomial systems can be sparse (when the number of terms in each equation is low), overdetermined (the number of the equations is larger than the number of variables), invariants by the action of some finite groups, multi-linear (each equation is linear w.r.t. to one block of variables) or more generally multihomogeneous. In each case, the ultimate goal is to identify large classes of problems whose theoretical/practical complexity drops and to propose in each case dedicated algorithms.

[0] J.-C. Faugère. *A new efficient algorithm for computing Gröbner bases without reduction to zero (F5).* In Proceedings of ISSAC '02, pages 75-83, New York, NY, USA, 2002. ACM.

3.3. Solving Systems over the Reals and Applications.

Participants: Mohab Safey El Din, Daniel Lazard, Elias Tsigaridas, Simone Naldi, Ivan Bannwarth.

We will develop algorithms for solving polynomial systems over complex/real numbers. Again, the goal is to extend significantly the range of reachable applications using algebraic techniques based on Gröbner bases and dedicated linear algebra routines. Targeted application domains are global optimization problems, stability of dynamical systems (e.g. arising in biology or in control theory) and theorem proving in computational geometry.

The following functionalities shall be requested by the end-users:
(i) deciding the emptiness of the real solution set of systems of polynomial equations and inequalities,
(ii) quantifier elimination over the reals or complex numbers,
(iii) answering connectivity queries for such real solution sets.
We will focus on these functionalities.

We will develop algorithms based on the so-called critical point method to tackle systems of equations and inequalities (problem *(i)*) . These techniques are based on solving 0-dimensional polynomial systems encoding "critical points" which are defined by the vanishing of minors of jacobian matrices (with polynomial entries). Since these systems are highly structured, the expected results of Objective 1 and 2 may allow us to obtain dramatic improvements in the computation of Gröbner bases of such polynomial systems. This will be the foundation of practically fast implementations (based on singly exponential algorithms) outperforming the current ones based on the historical Cylindrical Algebraic Decomposition (CAD) algorithm (whose complexity is doubly exponential in the number of variables). We will also develop algorithms and implementations that allow us to analyze, at least locally, the topology of solution sets in some specific situations. A long-term goal is obviously to obtain an analysis of the global topology.

3.4. Low level implementation and Dedicated Algebraic Computation and Linear Algebra.

Participants: Jean-Charles Faugère, Christian Eder, Elias Tsigaridas.

Here, the primary objective is to focus on *dedicated* algorithms and software for the linear algebra steps in Gröbner bases computations and for problems arising in Number Theory. As explained above, linear algebra is a key step in the process of computing efficiently Gröbner bases. It is then natural to develop specific linear algebra algorithms and implementations to further strengthen the existing software. Conversely, Gröbner bases computation is often a key ingredient in higher level algorithms from Algebraic Number Theory. In these cases, the algebraic problems are very particular and specific. Hence dedicated Gröbner bases algorithms and implementations would provide a better efficiency.

Dedicated linear algebra tools. FGB is an efficient library for Gröbner bases computations which can be used, for instance, via MAPLE. However, the library is sequential. A goal of the project is to extend its efficiency to new trend parallel architectures such as clusters of multi-processor systems in order to tackle a broader class of problems for several applications. Consequently, our first aim is to provide a durable, long term software solution, which will be the successor of the existing FGB library. To achieve this goal, we will first develop a high performance linear algebra package (under the LGPL license). This could be organized in the form of a collaborative project between the members of the team. The objective is not to develop a general library similar to the LINBOX project but to propose a dedicated linear algebra package taking into account the specific properties of the matrices generated by the Gröbner bases algorithms. Indeed these matrices are sparse (the actual sparsity depends strongly on the application), almost block triangular and not necessarily of full rank. Moreover, most of the pivots are known at the beginning of the computation. In practice, such matrices are huge (more than 10^6 columns) but taking into account their shape may allow us to speed up the computations by one or several orders of magnitude. A variant of a Gaussian elimination algorithm together with a corresponding C implementation has been presented. The main peculiarity is the order in which the operations are performed. This will be the kernel of the new linear algebra library that will be developed.

Fast linear algebra packages would also benefit to the transformation of a Gröbner basis of a zero–dimensional ideal with respect to a given monomial ordering into a Gröbner basis with respect to another ordering. In the generic case at least, the change of ordering is equivalent to the computation of the minimal polynomial of a so-called multiplication matrix. By taking into account the sparsity of this matrix, the computation of the Gröbner basis can be done more efficiently using variant of the Wiedemann algorithm. Hence, our goal is also to obtain a dedicated high performance library for transforming (i.e. change ordering) Gröbner bases.

Dedicated algebraic tools for Algebraic Number Theory. Recent results in Algebraic Number Theory tend to show that the computation of Gröbner basis is a key step toward the resolution of difficult problems in this domain [0]. Using existing resolution methods is simply not enough to solve relevant problems. The main algorithmic look to overcome is to adapt the Gröbner basis computation step to the specific problems. Typically, problems coming from Algebraic Number Theory usually have a lot of symmetries or the input systems are very structured. This is the case in particular for problems coming from the algorithmic theory of Abelian varieties over finite fields [0] where the objects are represented by polynomial system and are endowed with intrinsic group actions. The main goal here is to provide dedicated algebraic resolution algorithms and implementations for solving such problems. We do not restrict our focus on problems in positive characteristic. For instance, tower of algebraic fields can be viewed as triangular sets; more generally, related problems (e.g. effective Galois theory) which can be represented by polynomial systems will receive our attention. This is motivated by the fact that, for example, computing small integer solutions of Diophantine polynomial systems in connection with Coppersmith's method would also gain in efficiency by using a dedicated Gröbner bases computations step.

3.5. Solving Systems in Finite Fields, Applications in Cryptology and Algebraic Number Theory.

Participants: Jean-Charles Faugère, Ludovic Perret, Guénaël Renault, Louise Huot, Frédéric Urvoy de Portzamparc, Rina Zeitoun, Jérémy Berthomieu.

Here, we focus on solving polynomial systems over finite fields (i.e. the discrete case) and the corresponding applications (Cryptology, Error Correcting Codes, ...). Obviously this objective can be seen as an application of the results of the two previous objectives. However, we would like to emphasize that it is also the source of new theoretical problems and practical challenges. We propose to develop a systematic use of *structured systems* in *algebraic cryptanalysis*.
(i) So far, breaking a cryptosystem using algebraic techniques could be summarized as modeling the problem by algebraic equations and then computing a, usually, time consuming Gröbner basis. A new trend in this field is to require a theoretical complexity analysis. This is needed to explain the behavior of the attack but also to help the designers of new cryptosystems to propose actual secure parameters.
(ii) To assess the security of several cryptosystems in symmetric cryptography (block ciphers, hash functions, ...), a major difficulty is the size of the systems involved for this type of attack. More specifically, the bottleneck is the size of the linear algebra problems generated during a Gröbner basis computation.

We propose to develop a systematic use of *structured systems* in *algebraic cryptanalysis*.

[0] P. Gaudry, *Index calculus for abelian varieties of small dimension and the elliptic curve discrete logarithm problem*, Journal of Symbolic Computation 44,12 (2009) pp. 1690-1702
[0] e.g. point counting, discrete logarithm, isogeny.

The first objective is to build on the recent breakthrough in attacking McEliece's cryptosystem: it is the first structural weakness observed on one of the oldest public key cryptosystem. We plan to develop a well founded framework for assessing the security of public key cryptosystems based on coding theory from the algebraic cryptanalysis point of view. The answer to this issue is strongly related to the complexity of solving bihomogeneous systems (of bidegree $(1, d)$). We also plan to use the recently gained understanding on the complexity of structured systems in other areas of cryptography. For instance, the MinRank problem – which can be modeled as an overdetermined system of bilinear equations – is at the heart of the structural attack proposed by Kipnis and Shamir against HFE (one of the most well known multivariate public cryptosystem). The same family of structured systems arises in the algebraic cryptanalysis of the Discrete Logarithmic Problem (DLP) over curves (defined over some finite fields). More precisely, some bilinear systems appear in the polynomial modeling the points decomposition problem. Moreover, in this context, a natural group action can also be used during the resolution of the considered polynomial system.

Dedicated tools for linear algebra problems generated during the Gröbner basis computation will be used in algebraic cryptanalysis. The promise of considerable algebraic computing power beyond the capability of any standard computer algebra system will enable us to attack various cryptosystems or at least to propose accurate secure parameters for several important cryptosystems. Dedicated linear tools are thus needed to tackle these problems. From a theoretical perspective, we plan to further improve the theoretical complexity of the hybrid method and to investigate the problem of solving polynomial systems with noise, i.e. some equations of the system are incorrect. The hybrid method is a specific method for solving polynomial systems over finite fields. The idea is to mix exhaustive search and Gröbner basis computation to take advantage of the over-determinacy of the resulting systems.

Polynomial system with noise is currently emerging as a problem of major interest in cryptography. This problem is a key to further develop new applications of algebraic techniques; typically in side-channel and statistical attacks. We also emphasize that recently a connection has been established between several classical lattice problems (such as the Shortest Vector Problem), polynomial system solving and polynomial systems with noise. The main issue is that there is no sound algorithmic and theoretical framework for solving polynomial systems with noise. The development of such framework is a long-term objective.

POSTALE Team

3. Research Program

3.1. Architectures and program optimization

In this research topic, we focus on optimizing resources in a systematic way for the programmer by addressing fundamental issues like optimizing communication and data layout, generating automatically optimized codes via Domain Specific Languages (DSL), and auto-tuning of computer systems.

3.1.1. Optimization techniques for data and energy

3.1.1.1. Scientific context

Among the main challenges encountered in the race towards performance for supercomputers are energy (consumption, power and heat dissipation) and the memory/communication wall. This research topic addresses more specialized code analysis and optimization techniques as well as algorithmic changes in order to meet these two criteria, both from an expert - meaning handmade code transformations - or automatic - meaning compile time or run time - point of view.

Memory/communication wall means that processor elementary clock cycle decreases more rapidly over years than data transfer whether vertically between memory-ies and CPU (memory access) or horizontally between processors (data transfer). Moreover current architectures include complex memory features such as deep memory hierarchies, shared caches between cores, data alignment constraints, distributed memories etc. As a result data communication and data layout are becoming the bottleneck to performance and most program transformations aim at organizing them carefully and possibly avoiding or minimizing them. Energy consumption is also a limitation for today's processor performance. Then the options are either to design processors that consume less energy or, at the software level, to design energy-saving compilers and algorithms.

In general, the memory and energy walls are tackled with the same kind of program transformations that consist of avoiding as much as possible data communication [158] but considering these issues separately offers a different perspective. In this research axis, we focus on data/memory and energy/power optimization that include handmade or automatic compiler, code and algorithm optimizations. The resulting tools are expected to be integrated in other Postale topics related to auto-tuning [93], code generation [83] or communication-avoiding algorithms [51], [112].

3.1.1.2. Activity description and recent achievements

3.1.1.2.1. Optimization for data:

Program data transformation - data layout, data transfers. Postale has been addressing these issues in the past ANR PetaQCD project described in [63], [64] and in the PhD thesis of Michael Kruse [113]. The latter describes handmade data layout optimizations for optimizing a 4D stencil computation taking into account the BlueGene Q features. It also presents the Molly software based on the LLVM (Low Level Virtual Machine) Polly optimizing compiler that automatically generates code for MPI data transfers (see Figure 1 that shows an example of code generating a decomposition of a stencil computation into 4 subdomains and how data are exchanged between subdomains).

Data layout is still a critical point that Postale will address. The DSL [83] approach allows us to consider data layout globally, providing then an opportunity to study aggressive layouts without transformation penalty. We will also seize this opportunity to investigate the data layout problem as a new dimension of the CollectiveMind [93] optimization topic.

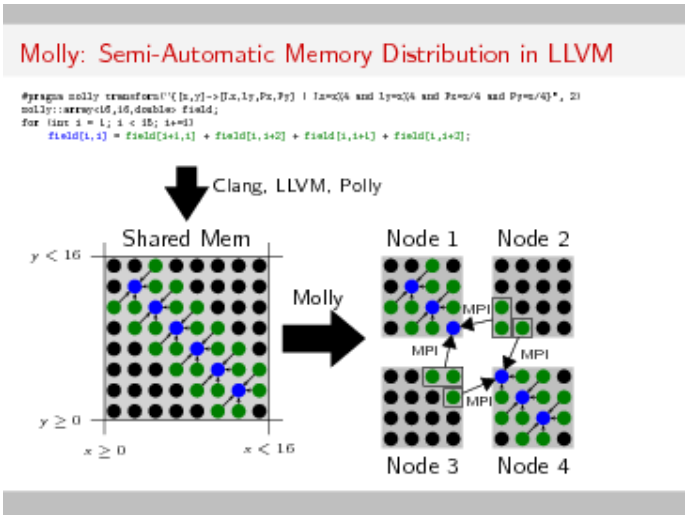


Figure 1. Automatic generation of subdomains using the Molly software.

Algorithm transformation - automating communication avoiding algorithms. This part is related to the Postale work on numerical algorithms. It originates from a research grant application elaborated with the former PetaQCD [64] team and the Inria Alpine project-team. One essential research direction consists of providing a set of high level optimizations that are generally out of reach from a traditional compiler approach. Among these optimizations, we consider communication-avoiding transformations and address the current open question of integrating these transformations in the polyhedral model in order to make them available in most software environments. Communication-avoiding algorithms improve parallelism and decrease communication requirements by ignoring some of dependency constraints at the frontiers of subdomains. Integrating communication-avoiding transformations is challenging first because these transformations change code semantics, which is unsual in program transformations, second because the validity of these transformations relies on numerical properties of the underlying transformed algorithms. This requires both compiler and algorithm skills since these transformations have important impact on the numerical stability and convergence of algorithms. Tools for the automatic generation of these transformed algorithms have two kinds of application. First, they accelerate the fastidious task of reprogramming for testing numerical properties. They may even be incorporated in an iterative tool for systematically evaluating these properties. Second, if these transformations are formalized we can consider generating different versions on line at run time, to adapt automatically algorithms to run time values [65]. In particular we plan to address s-steps algorithms [133] in iterative methods as these program transformations are similar to loop unrolling and ghosting (inverse of loop peeling). These are aggressive transformations and special preconditioning is needed in order to ensure convergence.

3.1.1.2.2. Optimizing energy:

In this topic there are two main research directions. The first one is about reversible computing based on the Landauer's conjecture that heat dissipation is produced by information erasing. The second one is on actual measurements of energy/power of program execution and on understanding which application features are the most likely to save or consume energy.

Regarding **reversible computing**, the Landauer's hypothesis - still in discussion among physicists - says that erasing one bit of information dissipates energy, independently from hardware. This implies that energy saving algorithms should avoid as much as possible erasing information: it should be possible to recover values of variables at any time in program execution. In a previous work we have analyzed the impact of

making computing DAG (Directed Acyclic Graphs) reversible [61]. We have also used reversible computing in register allocation by enabling value rematerialization also by reverse computing [62]. We are now working on characterizing algorithms by the amount of input and output data that have to be added to make algorithms reversible. We also plan to analyze mixed precision numerical algorithms [50] from this perspective.

Another research direction concerns **energy and power profiling and optimizing**. Understanding and monitoring precise energetic behavior of current programs is still a not easy task for the programmer or the compiler. One can measure it with wattmeters, or perform processor simulations or use hardware counters or sensors, or approximate it by the number of data that are communicated [159]. Especially on supercomputers or cloud framework it might be impossible to get this information. Besides making experiments on energy and power profiling [128], this research axis also includes the analysis of programming features that are the key parameters for saving energy. The ultimate goal is to have a cost model that describes the program energetic behavior of programs for the programmer or compiler being able to control it. One obvious key parameter is the count of memory accesses but one can also think of regularity features such as constant strides memory access, whether the code is statically or dynamically controlled, regularity/predictability conditional branchs. We have already performed this kind of analysis in the context of value prediction techniques where we designed entropy based criteria for estimating the predictibility of the sequence of values of some variables [129].

3.1.1.3. Research tracks for the 4 next years

Short term objectives are related to handmade or semi-automatic profiling and optimization of current scientific or image processing challenging applications. This gives a very good insight and expertise over state of the art applications and architectures. This know-how can be exploited under the form of libraries. This includes performance profiling, analysis of the energetic behavior of applications, and finding hot spots and focus optimization on these parts. This also implies to implement new numerical algorithms such as the communication-avoiding algorithms. Mid term objectives are to go forward to the automatization or semi-automatization of these techniques. Long term objectives are to understand the precise relationship between physics and computation both in programs as in reversible computing and in algorithms like in algorithmic thermodynamics [60]. The path is to define a notion of energetic complexity, which we intend to do it with the Galac team at Laboratoire de Recherche en Informatique.

3.1.2. Generative programming for new parallel architectures

3.1.2.1. Scientific context

Design, development and maintenance of high-performance scientific code is becoming one of the main issue of scientific computing. As hardware is becoming more complex and programming tools and models are proposed to satisfy constantly evolving applications, gathering expertise in both any scientific field and parallel programming is a daunting task. The natural conclusion is then to provide software design tools such that non-experts in computer science are able to produce non-trivial yet efficient codes on modern hardware architectures at their disposal. These tools can be divided in two types:

- **Compilers**. Compilers can be designed to either automatically derive parallel version of sequential codes or to support specific annotations to do so. Various successful examples include ISPC [137], SPADE [167] or GCC and its support for polyhedral compilation [140]. By offloading these tasks to compilers, the performance of the resulting codes is free of any overhead and the amount of user input is minimized. However, the scope and applicability of these techniques are fragile and can be hindered by complex code flow, inadequate data types or the use of high level languages features.
- **Libraries**. The unability of compilers to handle complex semantic is often mitigated by the design of libraries. Libraries can expose an arbitrary high level of abstraction through abstract data types and functions operating on them. User code is then expressed as a combination of function calls over instances of these data types. Different level of abstraction for parallel systems are available ranging from linear algebra [42], [109], image processing [70] to graph algorithms [153]. The main limitation of this approach is the lack of inter-procedural optimizations and the inherent divergence in API among vendors and targeted systems.

One emerging solution is to combine aspects of both solutions by designing systems which are able to provide abstraction and performance. One such approach is the design and development of **Domain Specific Languages** (or DSL) and more precisely, **Domain Specific Embedded Languages** (DSEL). DSLs [154] are non-general purpose, declarative language that simplify development by allowing users to express "the problem to solve" instead of "how to solve it". Actual code generation is then left to a proper compiler, interpreter or code generator that use high-level abstraction analysis and potential knowledge about target hardware to ensure performance. SCALA – and more precisely the FORGE tool [156] – is one of the most successful attempt at applying such techniques to parallel programming. DSELs differ from regular DSLs in the fact that they exist as a subset of an existing general purpose language. Often implemented as **Active Libraries** [166], they perform high-level optimizations based on a semantic analysis of the code before any real compilation process.

3.1.2.2. Activity description and recent achievements

In this research, we investigate the impact and applicability of software design methods based on DSELs to parallel programming and we study the portability and forward scalability of such programs. To do so, we investigate **Generative Programming** [76] applied to parallel programming.
Generative Programming is based on the hypothesis that any complex software system can be split into a list of interchangeable components (with clearly identified tasks) and a series of generators that combine components by following rules derived from an a priori domain specific analysis. In particular, we want to show that integrating the architectural support as another generative component of the set of tools leads to a better performance and an easier development on embedded or custom architecture targets (see Figure 2).

The application of Generative Programming allows us to build active libraries that can be easily re-targeted, optimized and deployed on a large selection of hardware systems. This is done by decoupling the abstract description of the DSEL from the description of hardware systems and the generation of hardware agnostic software components.
Current applications of this methodology include:

- BOOST.SIMD [84] is a C++ library for portable SIMD computations. It uses architecture aware generative programming to generate zero-overhead SIMD code on a large selection of platforms (from SSE to AVX2, Xeon Phi, PowerPC and ARM). Its interface is made so it is totally integrated into modern C++ design strategy based on the use of generic code and calls to the standard template libraries. In most cases, BOOST.SIMD delivers performance on the par with hand written SIMD code or with autovectorizers.
- NT^2 [83], [89] is a C++ library which implements a DSEL similar to MATLAB while providing automatic parallelization on SIMD systems, multicores and GPGPUs. NT^2 uses the high level of abstraction brought by the MATLAB API to detect, analyze and generate efficient loop nests taking care of every level of parallel hardware available. NT^2 eases the design of scientific computing application prototypes while delivering a significant percentage of the peak performance.

Our work uses a methodogy similar to SCALA [134], and more specifically, the DeLITE [157] toolset. Both approach rely on extracting high level, domain specific information from user code to optimize HPC applications. If our approach tries to maximize the use of compile-time optimization, DeLITE uses a runtime approach due to its reliance on the JAVA language.
In terms of libraries, various existing Scientific Computing library in C++ are actually available. The three most used are Armadillo [152], which shares a MATLAB-like API with our work, Blaze [69] which supports a similar cost based system for optimizing code and Eigen [100]. Our main feature compared to these solutions is the fact that hardware support is built-in the library core instead of beign tacked on the existing library, thus allowing us to support a larger amount of hardware.

3.1.2.3. Research tracks for the 4 next years

At short term, research and development on BOOST.SIMD and NT^2 will explore the applicability of our code generation methodology on distributed system, accelerators and heterogeneous systems. Large system support like Blue Gene/Q and other similar super-computer setup has been started.

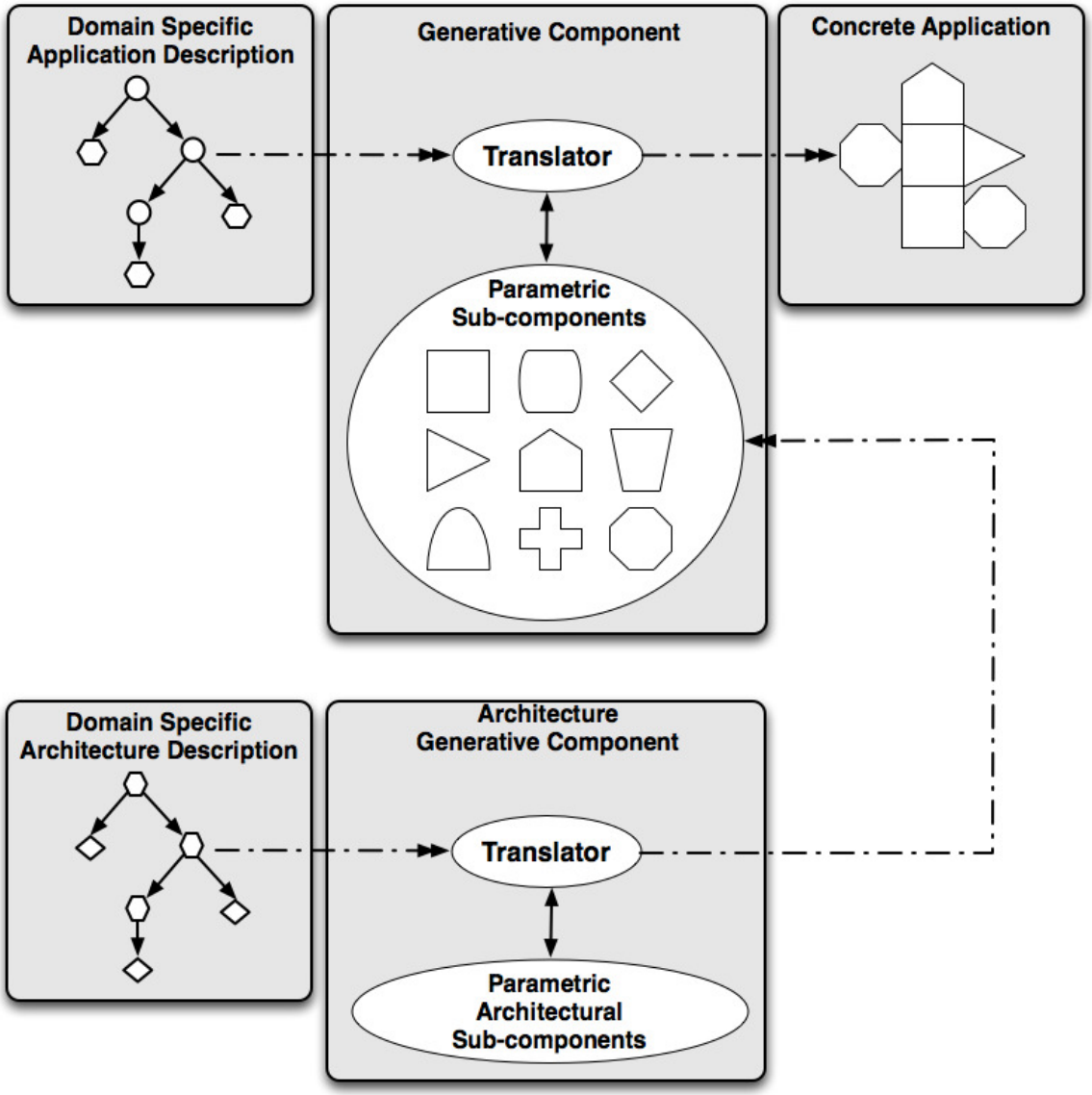


Figure 2. Principles of Architecture Aware Generative Programming

Another axis of research is to apply generative programing to other scientific domain and to propose other domain specific tools using efficient code generators. Such a work has been started to explore the impact of generative programming on the design of portable linear algebra algorithms with an going PhD thesis on automatic generation of linear algebra software.
A mid-term objective is to bridge the gap with the Data Analytics community in order to both extract new expertise on how to make Big Data related issues scalable on modern HPC hardware and to provide tools for Data Analytics practitioners based on this collaboration.
On a larger scope, the implication of our methodology on language design will be explored. First by proposing evolution to C++ (as for example with our SIMD proposal [85]) so that generative programming can become a first class citizen in the language itself. Second by exploring how this methodology can be extended to other languages [99] or to other runtime systems including Cloud computing systems and JIT support. Application to other performance metric like power consumption is also planned [171].

3.1.3. Systematizing and automating program optimization

3.1.3.1. Scientific context

Delivering faster, more power efficient and reliable computer systems is vital for our society to continue innovation in science and technology. However, program optimization and hardware co-design became excessively time consuming, costly and error prone due to an enormous number of available design and optimization choices, and complex interactions between all software and hardware components. Worse, multiple characteristics have to be always balanced at the same time including execution time, power consumption, code size, memory utilization, compilation time, communication costs and reliability using a growing number of incompatible tools and techniques with many ad-hoc and intuition based heuristics. As a result, nearly peak performance of the new systems is often achieved only for a few previously optimized and not necessarily representative benchmarks while leaving most of the real user applications severely underperforming. Therefore, users are often forced to resort to a tedious and often non-systematic optimization of their programs for each new architecture. This, in turn, leads to an enormous waste of time, expensive computing resources and energy, dramatically increases development costs and time-to-market for new products and slows down innovation [41], [39], [46], [80].

3.1.3.2. Activity description and recent achievements

For the european project MILEPOST (2006-2009) [40], we, for the first time to our knowledge, attempted to address above challenges in practice with several academic and industrial partners including IBM, CAPS, ARC (now Synopsys) and the University of Edinburgh by combining automatic program optimization and tuning, machine learning and a public repository of experimental results. As a part of the project, we established a non-profit cTuning association (cTuning.org) that persuaded the community to voluntarily support our open source tools and repository while sharing benchmarks, data sets, tools and machine learning models even after the project. This approach, highly prised by the European Commission, Inria and the international community, helped us to substitute and automatically learn best compiler optimization heuristics by crowdsourcing auto-tuning (processing a large amount of performance statistics or "big data" collected from many users to classify application and build predictive models) [40], [91], [92]. However, it also exposed even more fundamental challenges including:

- Lack of common, large and diverse benchmarks and data sets needed to build statistically meaningful predictive models;
- Lack of common experimental methodology and unified ways to preserve, systematize and share our growing optimization knowledge and research material from the community including benchmarks, data sets, tools, tuning plugins, predictive models and optimization results;
- Problem with continuously changing, “black box” and complex software and hardware stack with many hardwired and hidden optimization choices and heuristics not well suited for auto-tuning and machine learning;
- Difficulty to reproduce performance results from the cTuning.org database submitted by the community due to a lack of full software and hardware dependencies;

- Difficulty to validate related auto-tuning and machine learning techniques from existing publications due to a lack of culture of sharing research artifacts with full experiment specifications along with publications in computer engineering.

As a result, we spent a considerable amount of our "research" time on re-engineering existing tools or developing new ones to support auto-tuning and learning. At the same time, we were trying to somehow assemble large and diverse experimental sets to make our research and experimentation on machine learning and data mining statistically meaningful. We spent even more time when struggling to reproduce existing machine learning-based optimization techniques from numerous publications. Worse, when we were ready to deliver auto-tuning solutions at the end of such tedious developments, experimentation and validation, we were already receiving new versions of compilers, third-party tools, libraries, operating systems and architectures. As a consequence, our developments and results were already potentially outdated even before being released while optimization problems considerably evolved.

We believe that these are major reasons why so many promising research techniques, tools and data sets for auto-tuning and machine learning in computer engineering have a life span of a PhD project, grant funding or publication preparation, and often vanish shortly after. Furthermore, we witness diminishing attractiveness of computer engineering often seen by students as "hacking" rather than systematic science. Many recent long-term research visions acknowledge these problems for computer engineering and many research groups search for "holy grail" auto-tuning solutions but no widely adopted solution has been found yet [39], [80].

3.1.3.3. Research tracks for the 4 next years

In this project, we will be evaluating the first, to our knowledge, alternative, orthogonal, interdisciplinary, community-based and big-data driven approach to address above problems. We are developing a knowledge management system for computer engineering (possibly based on GPL-licensed cTuning and BSD-licensed Collective Mind) to preserve and share through the Internet the whole experimental (optimization) setups with all related artifacts and exposed meta-description in a unified way including behavior characteristics (execution time, code size, compilation time, power consumption, reliability, costs), semantic and dynamic features, design and optimization choices, and a system state together with all software and hardware dependencies besides just performance data. Such approach allows community to consider analysis, design and optimization of computer systems as a unified, formalized and big data problem while taking advantage of mature R&D methodologies from physics, biology and AI.

During this project, we will gradually structure, systematize, describe and share all research material in computer engineering including tools, benchmarks, data sets, search strategies and machine learning models. Researchers can later take advantage of shared components to collaboratively prototype, evaluate and improve various auto-tuning techniques while reusing all shared artifacts just like LEGO™pieces, and applying machine learning and data mining techniques to find meaningful relations between all shared material. It can also help crowdsource long tuning and learning process including classification and model building among many participants.

At the same time, any unexpected program behavior or model mispredictions can now be exposed to the community through unified web-services for collaborative analysis, explanation and solving. This, in turn, enables reproducibility of experimental results naturally and as a side effect rather than being enforced - interdisciplinary community needs to gradually find and add missing software and hardware dependencies to the Collective Mind (fixing processor frequency, pinning code to specific cores to avoid contentions) or improve analysis and predictive models (statistical normality tests for multiple experiments) whenever abnormal behavior is detected.

We hope that our approach will eventually help the community collaboratively evaluate and derive the most effective optimization strategies. It should also eventually help the community collaboratively learn complex behavior of all existing computer systems using top-down methodology originating from physics. At the same time, continuously collected and systematized knowledge ("big data") should allow community make quick and scientifically motivated advice about how to design and optimize the future heterogeneous HPC systems (particularly on our way towards extreme scale computing) as conceptually shown in Figure 3 .

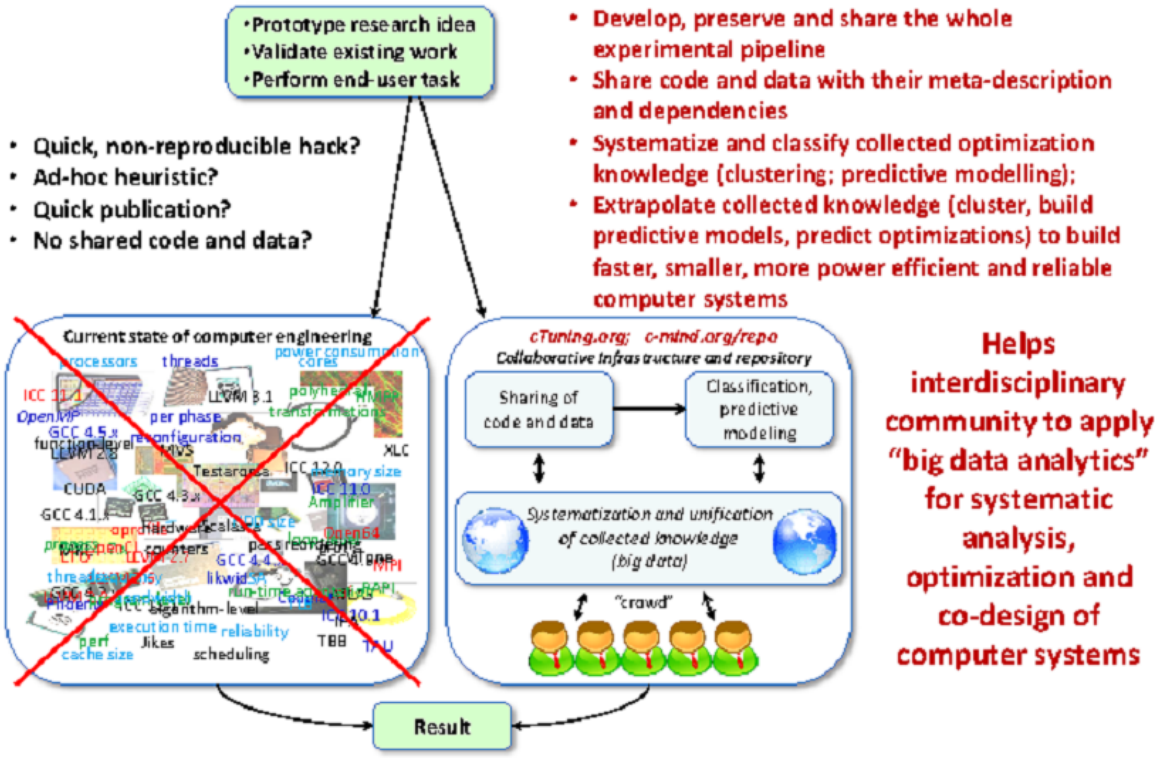


Figure 3. *Considering program optimization and run-time adaptation as a "big data problem"*

Similar systematization, formalization and big data analytics already revolutionized biology, machine learning, robotics, AI, and other important scientific fields in the past decade. Our approach also started revolutionizing computer engineering making it more a science rather than non-systematic hacking. It helps us effectively deal with the rising complexity of computer systems while focusing on improving classification and predictive models of computer systems' behavior, and collaboratively find missing features (possibly using new deep learning algorithms and even unsupervised learning [106], [126]) to improve optimization predictions, rather than constantly reinventing techniques for each new program, architecture and environment.

Our approach is strongly supported by a recent Vinton G. Cerf's vision for computer engineering [73] as well as our existing technology, repository of knowledge and experience, and a growing community [91], [92], [93]. Even more importantly, our approach already helped to promote reproducible research and initiate a new publication model in computer engineering supported by ACM SIGPLAN where all experimental results and related research artifacts with their meta-description and dependencies are continuously shared along with publications to be validated and improved by the community [90].

3.2. High-level HPC libraries and applications

In this research topic, we focus on developing optimized algorithms and software for high-performance scientific computing and image processing.

3.2.1. Taking advantage of heterogeneous parallel architectures

3.2.1.1. Activity description

In recent years and as observed in the latest trends from the Top 500 list [0], heterogeneous computing combining manycore systems with accelerators such as Graphics Processing Units (GPU) or Intel Xeon Phi coprocessors has become a *de facto* standard in high performance computing. At the same time, data movements between memory hierarchies and/or between processors have become a major bottleneck for most numerical algorithms. The main goal of this topic is to investigate new approaches to develop linear algebra algorithms and software for heterogeneous architectures [56], [164], with also the objective of contributing to public domain numerical linear algebra libraries (e.g., MAGMA [0]).
Our activity in the field consists of designing algorithms that minimize the cost of communication and optimize data locality in numerical linear algebra solvers. When combining different architectures, these algorithms should be properly "hybridized". This means that the workload should be balanced throughout the execution, and the work scheduling/mapping should ensure matching of architectural features to algorithmic requirements.

In our effort to minimize communication, an example concerns the solution of general linear systems (via LU factorization) where the main objective is to reduce the communication overhead due to pivoting. We developed several algorithms to achieve this objective for hybrid CPU/GPU platforms. In one of them the panel factorization is performed using a communication-avoiding pivoting heuristic [97] while the update of the trailing submatrix is performed by the GPU [51]. In another algorithm, we use a random preconditioning (see also Section 3.2.2) of the original matrix to avoid pivoting [54]. Performance comparisons and tests on accuracy showed that these solvers are effective on current hybrid multicore-GPU parallel machines. These hybrid solvers will be integrated in a next release of the MAGMA library.

Another issue is related to the impact of non-uniform memory accesses (NUMA) on the solution of HPC applications. For dense linear systems, we illustrated how an appropriate placement of the threads and memory on a NUMA architecture can improve the performance of the panel factorization and consequently accelerate the global LU factorization [148], when compared to the hybrid multicore/GPU LU algorithm as it is implemented in the public domain library MAGMA.

[0] http://www.top500.org/
[0] Matrix Algebra on GPU and Multicore Architectures, http://icl.cs.utk.edu/magma/

3.2.1.2. Research tracks for the 4 next years

3.2.1.2.1. Towards automatic generation of dense linear solvers:

In an ongoing research, we investigate a generic description of the linear system to be solved in order to exploit numerical and structural properties of matrices to get fast and accurate solutions with respect to a specific type of problem. Information about targeted architectures and resources available will be also taken into account so that the most appropriate routines are used or generated. An application of this generative approach is the possibility of prototyping new algorithms or new implementations of existing algorithms for various hardware.

A track for generating efficient code is to develop new functionalities in the C++ library NT^2 [89] which is developed in the Postale team. This approach will enable us to generate optimized code that support current processor facilities (OpenMP and TBB support for multicores, SIMD extensions...) and accelerators (GPU, Intel Xeon Phi) starting from an API (Application Programming Interface) similar to Matlab. By analyzing the properties of the linear algebra domain that can be extracted from numerical libraries and combining them with architectural features, we have started to apply the generic approach mentioned in Section 3.1.2 to solve dense linear systems on various architectures including CPU and GPU. As an application, we plan to develop a new software that can run either on CPU or GPU to solve least squares problems based on semi-normal equations in mixed precision [50] since, to our knowledge, such a solver cannot be found in current public domain libraries (Sca)LAPACK [43], [68], PLASMA [165] and MAGMA [52]. This solver aims at attaining a performance that corresponds to what state-of-the-art codes achieve using mixed precision algorithms.

3.2.1.2.2. Communication avoiding algorithms for heterogeneous platforms:

In previous work, we focused on the LU decomposition with respect to two directions that are numerical stability and communication issue. This research work has lead to the development of a new algorithm for the LU decomposition, referred to as LU_PRRP: LU with panel rank revealing pivoting [112]. This algorithm uses a new pivoting strategy based on strong rank revealing QR factorization [98]. We also design a communication avoiding version of LU_PRRP, referred to as CALU_PRRP, which aims at overcoming the communication bottleneck during the panel factorization if we consider a parallel version of LU_PRRP. Thus CALU_PRRP is asymptotically optimal in terms of both bandwidth and latency. Moreover, it is more stable than the communication avoiding LU factorization based on Gaussian elimination with partial pivoting in terms of growth factor upper bound [78].

Due to the huge number and the heterogeneity of computing units in future exascale platforms, it is crucial for numerical algorithms to exhibit more parallelism and pipelining. It is thus important to study the critical paths of these algorithms, the task decomposition and the task granularity as well as the scheduling techniques in order to take advantage of the potential of the available platforms. Our goal here is to adapt our new algorithm CALU_PRRP to be scalable and efficient on heterogeneous platforms making use of the available accelerators and coprocessors similarly to what was achieved in [51].

3.2.1.2.3. Application to numerical fluid mechanics:

In an ongoing PhD thesis [168], [169], we apply hybrid programming techniques to develop a solver for the incompressible Navier-Stokes equations with constant coefficients, discretized by the finite difference method. In this application, we focus on solving large sparse linear systems coming from the discretization of Helmholtz and Poisson equations using direct methods that represent the major part of the computational time for solving the Navier-Stokes equations which describe a large class of fluid flows. In the future, our effort in the field will concern how to apply hybrid programming techniques to solvers based on iterative methods. A major task will consist of developing efficient kernels and choosing appropriate preconditioners. An important aspect is also the use of advanced scheduling techniques to minimize the number of synchronizations during the execution. The algorithms developed during this research activity will be validated on physical data provided by the physicists either form the academic world (e.g., LIMSI/University Paris-Sud [0] or industrial partners (e.g., EDF, ONERA). This research is currently performed in the framework of the CALIFHA project [0] and will be continued in an industrial contract with EDF R&D (starting October 2014).

[0] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, http://www.limsi.fr/

[0] CALculations of Incompressible Fluids on Heterogeneous, funded by Région Île-de-France and Digitéo (http://www.digiteo.fr)

3.2.2. Randomized algorithms in HPC applications

Activity description

Randomized algorithms are becoming very attractive in high-performance computing applications since they are able to outperform deterministic methods while still providing accurate results. Recent advances in the field include for instance random sampling algorithms [47], low-rank matrix approximation [130], or general matrix decompositions [101].

Our research in this domain consists of developing fast algorithms for linear algebra solvers which are at the heart of many HPC physical applications. In recent works, we designed randomized algorithms [54], [66] based on random butterfly transformations (RBT) [135] that can be applied to accelerate the solution of general or symmetric indefinite (dense) linear systems for multicore [49] or distributed architectures [48]. These randomized solvers have the advantage of reducing the amount of communication in dense factorizations by removing completely the pivoting phase which inhibits performance in Gaussian Elimination.

We also studied methods and software to assess the numerical quality of the solution computed in HPC applications. The objective is to compute quantities that provide us with information about the numerical quality of the computed solution in an acceptable time, at least significantly cheaper than the cost for the solution itself (typically a statistical estimation should require $\mathcal{O}(n^2)$ flops while the solution of a linear system involves at least $\mathcal{O}(n^3)$ flops, where n is the problem size). In particular, we recently applied in [58] statistical techniques based on the small sample theory [111] to estimate the condition number of linear system/linear least squares solvers [45], [53], [57]. This approach reduces significantly the number of arithmetic operations in estimating condition numbers. Whether designing fast solvers or error analysis tools, our ultimate goal is to integrate the resulting software into HPC libraries so that these routines will be available for physicists. The targeted architectures are multicore systems possibly accelerated with GPUs or Intel Xeon Phi coprocessors.

This research activity benefits from the Inria associate-team program, through the **associate-team R-LAS**[0], created in 2014 between Inria Saclay/Postale team and University of Tennessee (Innovative Computing Laboratory) in the area of randomized algorithms and software for numerical linear algebra. This project is funded from 2014 to 2016 and is lead jointly by Marc Baboulin (Inria/University Paris-Sud) and Jack Dongarra (University of Tennessee).

Research tracks for the 4 next years

3.2.2.1. Extension of random butterfly transformations to sparse matrices:

We recently illustrated how randomization via RBT can accelerate the solution of dense linear systems on multicore architectures possibly accelerated by GPUs. We recently started to extend this method to sparse linear systems arising from the discretization of partial differential equations in many physical applications. However, a major difficulty comes from the possible fill-in introduced by RBT. One of our first task consists of performing experiments on a collection of sparse matrices to evaluate the fill-in depending on the number of recursions in the algorithm. In a recent work [59], we investigated the possibility of using another form of RBT (one-side RBT instead of two-sided) in order to minimize the fill-in and we obtain promising preliminary results (Figure 4 shows that the fill-in is significantly reduced when using one-side RBT).

Another track of research is related to iterative methods for solving large sparse linear systems, and more particularly preconditioned Krylov subspace methods implemented in the solver ARMS (Algebraic Recursive Multilevel Solver (pARMS for its parallel distributed version). In this solver, our goal is to find the last level of preconditioning and then replace the original ILU factorization by our RBT preprocessing. A PhD thesis (supervised by Marc Baboulin) started in October 2014 on using randomization techniques like RBT for sparse linear systems.

[0] Randomized Linear Algebra Software, https://www.lri.fr/~baboulin/presentation_r-las.html/

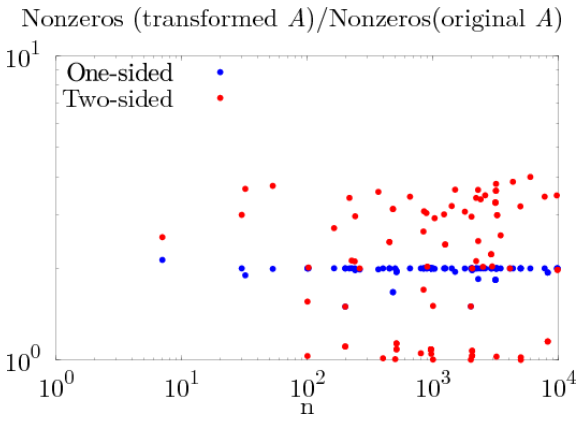


Figure 4. Evaluation of fill-in for one-sided RBT (90 matrices sorted by size).

3.2.2.2. Randomized algorithms on large clusters of multicore:

A major challenge for the randomized algorithms that we develop is to be able to solve very large problems arising in real-world physical simulations. As a matter of fact, large-scale linear algebra solvers from standard parallel distributed libraries like ScaLAPACK often suffer from expensive inter-node communication costs. An important requirement is to be able to schedule these algorithms dynamically on highly distributed and heterogeneous parallel systems [110]. In particular we point out that even though randomizing linear systems removes the communication due to pivoting, applying recursive butterflies also requires communication, especially if we use multiple nodes to perform the randomization. Our objective is to minimize this communication in the tiled agorithms and to use a runtime that enforces a strict data locality scheduling strategy [48]. A state of the art of possible runtime systems and how they can be combined with our randomized solvers will be established. Regarding the application of such solver, a collaboration with Pr Tetsuya Sakurai (University of Tsukuba, Japan) and Pr Jose Roman (Universitat Politècnica de València, Spain) will start in December 2014 to apply RBT to large linear systems encountered in contour integral eigensolver (CISS) [108]. Optimal tuning of the code will be obtained using holistic approach developed in the Postale team [93].

3.2.2.3. Extension of statistical estimation techniques to eigenvalue and singular value problems:

The extension of statistical condition estimation techniques can be carried out for eigenvalue/singular value calculations associated with nonsymmetric and symmetric matrices arising in, for example, optimization problems. In all cases, numerical sensitivity of the model parameters is of utmost concern and will guide the choice of estimation techniques. The important class of componentwise relative perturbations can be easily handled for a general matrix [111]. A significant outcome of the research will be the creation of high-quality open-source implementations of the algorithms developed in the project, similarly to the equivalent work for least squares problems [55]. To maximize its dissemination and impact, the software will be designed to be extensible, portable, and customizable.

3.2.2.4. Random orthogonal matrices:

Random orthogonal matrices have a wide variety of applications. They are used in the generation of various kinds of random matrices and random matrix polynomials [67], [77], [79], [105]. They are also used in some finance and statistics applications. For example the random orthogonal matrix (ROM) simulation [127] method uses random orthogonal matrices to generate multivariate random samples with the same mean and covariance as an observed sample.

The natural distribution over the space of orthogonal matrices is the Haar distribution. One way to generate a random orthogonal matrix from the Haar distribution is to generate a random matrix A with elements from the standard normal distribution and compute its QR factorization $A = QR$, where R is chosen to have nonnegative diagonal elements; the orthogonal factor Q is then the required matrix [104].

Stewart [155] developed a more efficient algorithm that directly generates an $n \times n$ orthogonal matrix from the Haar distribution as a product of Householder transformations built from Householder vectors of dimensions $1, 2, \cdots, n-1$ chosen from the standard normal distribution. Our objective is to design an algorithm that significantly reduces the computational cost of Stewart's algorithm by relaxing the property that Q is exactly Haar distributed. We also aim at extending the use of random orthogonal matrices to other randomized algorithms.

3.2.3. *Embedded high-performance systems & computer vision*

Scientific context

High-performance embedded systems & computer vision address the design of efficient algorithms for parallel architectures that deal with image processing and computer vision. Such systems must enforce realtime execution constraint (typically 25 frames per second) and power consumption constraint. If no COTS (*Component On The Shelf*) architecture (e.g., SIMD multicore processor, GPU, Intel Xeon Phi, DSP) satisfy the constraints, then we have to develop a specialized one.

A more and more important aspect when designing an embedded system is the tradeoff between speed (and power consumption) and numerical accuracy (and stability). Such a tradeoff leads to 16-bit computation (and storage) and to the design of less accurate algorithms. For example, the final accuracy for stabilizing an image is $10-1$ pixel, which is far from the maximum accuracy of (10^{-7}) available using the 32-bit IEEE format.

3.2.3.1. Activity description and recent achievements

Concerning image processing, our efforts concern the redesign of data-dependent algorithms for parallel architectures. A representative example of such an algorithm is the connected-component labeling (CCL) algorithm [147] which is used in industrial or medical imaging and classical computer vision like optical character recognition. As far as we know our algorithm (*Light Speed Labeling*) [71], [72] still outperforms other existing CCL algorithms [96], [103], [160] (the first versions of our algorithm appeared in 2009 [119], [120]).

Concerning computer vision (smart camera, autonomous robot, aerial drone), we developed in collaboration with LIMSI [0] two applications that run in realtime on embedded parallel systems [121], [146] with some accuracy tradeoffs. The first one is based on mean shift tracking [94], [95] and the second one relies on covariance matching and tracking [143], [144], [145].

These applications are used in video-surveillance: they perform motion detection [118], motion analysis [161], [162], motion estimation and multi-target tracking. Depending on the image nature and size, some algorithmic transforms (integral image, cumulative differential sum) can be applied and combined with hybrid arithmetic (16-bit / 32-bit / 64-bit). Finally, to increase the algorithm robustness color, space optimization is also used [122].

Usually one tries to convert 64-bit computations into 32-bit. But sometimes 16-bit floating point arithmetic is sufficient. As 16-bit numbers are now normalized by IEEE (754-2008) and are available in COTS processors like GPU and GPP (AVX2 for storage in memory and conversion into 32-bit numbers), we can run such kind of code on COTS processors or we can design specialized architectures like FPGA (*Field-Programmable gate array*) and ASIC (*Application-specific integrated circuit*) to be more efficient. This approach is complementary to that of [131] which converts 32-bit floating point signal processing operators into fixed-point ones.

[0] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

By extension to computer vision, we also address *interactive sensing HPC applications*. One CEA thesis funded by CEA and co-supervised by Lionel Lacassagne addresses the parallelization of Non Destructive Testing applications on COTS processors (super-charged workstation with GPUs and Intel Xeon Phi manycore processor). This PhD thesis deals with irregular computations with sparse-addressing and load-balacing problems. It also deals with floating point accuracy, by finding roots of polynomials using Newton and Laguerre algorithms. Depending on the configuration, 64-bit is required, but sometime 32-bit computations are sufficient with respect to the physics. As the second application focuses on interactive sensing, one has to add a second level of tradeoff for physical sampling accuracy and the sensor displacement [123], [124], [125], [141], [142].
In order to achieve realtime execution on the targeted architectures, we develop *High Level Transforms* (HLT) that are algorithmic transforms for memory layout and function re-organization. We show on a representative algorithm [102] in the image processing area that a fully parallelized code (SIMD+OpenMP) can be accelerated by a factor $\times 80$ on a multicore processor [115]. A CIFRE thesis (defended in 2014) funded by ST Microelectronics and supervised by Lionel Lacassagne has led to the design of very efficient implementations into an ASIC thanks to HLT. We show that the power consumption can be reduced by a factor 10 [170], [171].
All these applications have led to the development of software libraries for image processing that are currently under registration at APP (Agence de Protection des Programmes): `myNRC 2.0`[0] and `covTrack`[0].

3.2.3.2. Future: system, image & arithmetic

Concerning image processing we are designing new versions of CCL algorithms. One version is for parallel architectures where graph merging and efficient transitive closure is a major issue for load balancing. For embedded systems, *time prediction* is as important as execution time, so a specialized version targets embedded processors like ARM processors and Texas Instrument VLIW DSP C6x.

We also plan to design algorithms that should be less data-sensitive (the execution time depends on the nature of the image: a structured image can be processed quickly whereas an unstructured image will require more time). These algorithms will be used in even more data-dependent algorithms like *hysteresis thresholding* for image binarization, *split-and-merge* [44], [114] for realtime image segmentation using the Horowitz-Pavlidis quad-tree decomposition [107]. Such an algorithm could be useful for accelerating image decomposition like *Fast Level Set Transform* algorithm [132].

Concerning Computer vision we will study 16-bit floating point arithmetic for image processing applications and linear algebra operators. Concerning image processing, we will focus on iterative algorithms like optical flow computation (for motion estimation and image stabilization). We will compare the efficiency (accuracy and speed) of 16-bit floating point [86], [117], [116], [139] with fixed-point arithmetic. Concerning linear algebra, we will study efficient implementation for very small matrix inversion (from 6×6 up to 16×16) for our covariance-tracking algorithm.

According to Nvidia (see Figure 5), the computation rate (Gflop/s) for ZGEMM (complex matrix-matrix multiplication with 64-bit precision – for small value of N – is linearly proportional to N. That means that, for a 6×6 matrix, we achieve around 6 Gflop/s on a Tesla M2090 (400 Gflop/s peak power). This represents 1.5 % of the peak power. For that reason, designing efficient parallel codes for embedded systems [74], [81], [82] is different and may be more complex than designing codes for classical HPC systems. Our `covTrack` software requires many hundreds of 6×6 matrix-matrix multiplications every frame.
Last point is to develop tools that help to automatically distribute or parallelize a code on an architecture code parallelization/distribution dealing with scientific computing [83], MPI [87] or image applications on the Cell processor [75], [88], [138], [149], [150], [151], [163].

[0] smart memory allocator and management for 2D and 3D image processing
[0] agile realtime multi-target tracking algorithm, co-developped with Michèle Gouiffès at LIMSI

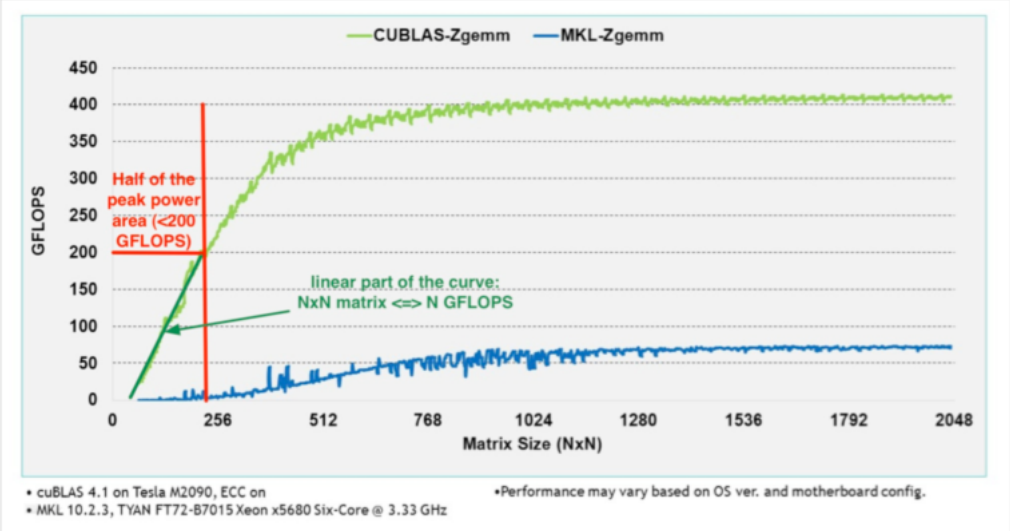


Figure 5. Nvidia cuBLAS performance versus Intel MKL: both have poor performance for small N

PRIVATICS Project-Team (section vide)

PROSECCO Project-Team

3. Research Program

3.1. Symbolic verification of cryptographic applications

Despite decades of experience, designing and implementing cryptographic applications remains dangerously error-prone, even for experts. This is partly because cryptographic security is an inherently hard problem, and partly because automated verification tools require carefully-crafted inputs and are not widely applicable. To take just the example of TLS, a widely-deployed and well-studied cryptographic protocol designed, implemented, and verified by security experts, the lack of a formal proof about all its details has regularly led to the discovery of major attacks (including several in 2014) on both the protocol and its implementations, after many years of unsuspecting use.

As a result, the automated verification for cryptographic applications is an active area of research, with a wide variety of tools being employed for verifying different kinds of applications.

In previous work, the we have developed the following three approaches:

- ProVerif: a symbolic prover for cryptographic protocol models
- Tookan: an attack-finder for PKCS#11 hardware security devices
- F7: a security typechecker for cryptographic applications written in F#

3.1.1. Verifying cryptographic protocols with ProVerif

Given a model of a cryptographic protocol, the problem is to verify that an active attacker, possibly with access to some cryptographic keys but unable to guess other secrets, cannot thwart security goals such as authentication and secrecy [86]; it has motivated a serious research effort on the formal analysis of cryptographic protocols, starting with [84] and eventually leading to effective verification tools, such as our tool ProVerif.

To use ProVerif, one encodes a protocol model in a formal language, called the applied pi-calculus, and ProVerif abstracts it to a set of generalized Horn clauses. This abstraction is a small approximation: it just ignores the number of repetitions of each action, so ProVerif is still very precise, more precise than, say, tree automata-based techniques. The price to pay for this precision is that ProVerif does not always terminate; however, it terminates in most cases in practice, and it always terminates on the interesting class of *tagged protocols* [81]. ProVerif also distinguishes itself from other tools by the variety of cryptographic primitives it can handle, defined by rewrite rules or by some equations, and the variety of security properties it can prove: secrecy [79], [71], correspondences (including authentication) [80], and observational equivalences [78]. Observational equivalence means that an adversary cannot distinguish two processes (protocols); equivalences can be used to formalize a wide range of properties, but they are particularly difficult to prove. Even if the class of equivalences that ProVerif can prove is limited to equivalences between processes that differ only by the terms they contain, these equivalences are useful in practice and ProVerif is the only tool that proves equivalences for an unbounded number of sessions.

Using ProVerif, it is now possible to verify large parts of industrial-strength protocols such as TLS [75], JFK [72], and Web Services Security [77]. against powerful adversaries that can run an unlimited number of protocol sessions, for strong security properties expressed as correspondence queries or equivalence assertions. ProVerif is used by many teams at the international level, and has been used in more 30 research papers (references available at http://proverif.inria.fr/proverif-users.html).

3.1.2. Verifying security APIs using Tookan

Security application programming interfaces (APIs) are interfaces that provide access to functionality while also enforcing a security policy, so that even if a malicious program makes calls to the interface, certain security properties will continue to hold. They are used, for example, by cryptographic devices such as smartcards and Hardware Security Modules (HSMs) to manage keys and provide access to cryptographic functions whilst keeping the keys secure. Like security protocols, their design is security critical and very difficult to get right. Hence formal techniques have been adapted from security protocols to security APIs.

The most widely used standard for cryptographic APIs is RSA PKCS#11, ubiquitous in devices from smartcards to HSMs. A 2003 paper highlighted possible flaws in PKCS#11 [82], results which were extended by formal analysis work using a Dolev-Yao style model of the standard [83]. However at this point it was not clear to what extent these flaws affected real commercial devices, since the standard is underspecified and can be implemented in many different ways. The Tookan tool, developed by Steel in collaboration with Bortolozzo, Centenaro and Focardi, was designed to address this problem. Tookan can reverse engineer the particular configuration of PKCS#11 used by a device under test by sending a carefully designed series of PKCS#11 commands and observing the return codes. These codes are used to instantiate a Dolev-Yao model of the device's API. This model can then be searched using a security protocol model checking tool to find attacks. If an attack is found, Tookan converts the trace from the model checker into the sequence of PKCS#11 queries needed to make the attack and executes the commands directly on the device. Results obtained by Tookan are remarkable: of 18 commercially available PKCS#11 devices tested, 10 were found to be susceptible to at least one attack.

3.1.3. Verifying cryptographic applications using F7 and F*

Verifying the implementation of a protocol has traditionally been considered much harder than verifying its model. This is mainly because implementations have to consider real-world details of the protocol, such as message formats, that models typically ignore. This leads to a situation that a protocol may have been proved secure in theory, but its implementation may be buggy and insecure. However, with recent advances in both program verification and symbolic protocol verification tools, it has become possible to verify fully functional protocol implementations in the symbolic model.

One approach is to extract a symbolic protocol model from an implementation and then verify the model, say, using ProVerif. This approach has been quite successful, yielding a verified implementation of TLS in F# [75]. However, the generated models are typically quite large and whole-program symbolic verification does not scale very well.

An alternate approach is to develop a verification method directly for implementation code, using well-known program verification techniques such as typechecking. F7 [73] is a refinement typechecker for F#, developed jointly at Microsoft Research Cambridge and Inria. It implements a dependent type-system that allows us to specify security assumptions and goals as first-order logic annotations directly inside the program. It has been used for the modular verification of large web services security protocol implementations [76]. F* [87] is an extension of F7 with higher-order kinds and a certifying typechecker. Both F7 and F* have a growing user community. The cryptographic protocol implementations verified using F7 and F* already represent the largest verified cryptographic applications to our knowledge.

3.2. Computational verification of cryptographic applications

Proofs done by cryptographers in the computational model are mostly manual. Our goal is to provide computer support to build or verify these proofs. In order to reach this goal, we have already designed the automatic tool CryptoVerif, which generates proofs by sequences of games. Much work is still needed in order to develop this approach, so that it is applicable to more protocols. We also plan to design and implement techniques for proving implementations of protocols secure in the computational model, by generating them from CryptoVerif specifications that have been proved secure, or by automatically extracting CryptoVerif models from implementations.

A different approach is to directly verify cryptographic applications in the computational model by typing. A recent work [85] shows how to use refinement typechecking in F7 to prove computational security for protocol implementations. In this method, henceforth referred to as computational F7, typechecking is used as the main step to justify a classic game-hopping proof of computational security. The correctness of this method is based on a probabilistic semantics of F# programs and crucially relies on uses of type abstraction and parametricity to establish strong security properties, such as indistinguishability.

In principle, the two approaches, typechecking and game-based proofs, are complementary. Understanding how to combine these approaches remains an open and active topic of research.

An alternative to direct computation proofs is to identify the cryptographic assumptions under which symbolic proofs, which are typically easier to derive automatically, can be mapped to computational proofs. This line of research is sometimes called computational soundness and the extent of its applicability to real-world cryptographic protocols is an active area of investigation.

3.3. Provably secure web applications

Web applications are fast becoming the dominant programming platform for new software, probably because they offer a quick and easy way for developers to deploy and sell their *app*s to a large number of customers. Third-party web-based apps for Facebook, Apple, and Google, already number in the hundreds of thousands and are likely to grow in number. Many of these applications store and manage private user data, such as health information, credit card data, and GPS locations. To protect this data, applications tend to use an ad hoc combination of cryptographic primitives and protocols. Since designing cryptographic applications is easy to get wrong even for experts, we believe this is an opportune moment to develop security libraries and verification techniques to help web application programmers.

As a typical example, consider commercial password managers, such as LastPass, RoboForm, and 1Password. They are implemented as browser-based web applications that, for a monthly fee, offer to store a user's passwords securely on the web and synchronize them across all of the user's computers and smartphones. The passwords are encrypted using a master password (known only to the user) and stored in the cloud. Hence, no-one except the user should ever be able to read her passwords. When the user visits a web page that has a login form, the password manager asks the user to decrypt her password for this website and automatically fills in the login form. Hence, the user no longer has to remember passwords (except her master password) and all her passwords are available on every computer she uses.

Password managers are available as browser extensions for mainstream browsers such as Firefox, Chrome, and Internet Explorer, and as downloadable apps for Android and Apple phones. So, seen as a distributed application, each password manager application consists of a web service (written in PHP or Java), some number of browser extensions (written in JavaScript), and some smartphone apps (written in Java or Objective C). Each of these components uses a different cryptographic library to encrypt and decrypt password data. How do we verify the correctness of all these components?

We propose three approaches. For client-side web applications and browser extensions written in JavaScript, we propose to build a static and dynamic program analysis framework to verify security invariants. To this end, we have developed two security-oriented type systems for JavaScript, Defensive JavaScript [74],[65] and TS* [62], and used them to guarantee security properties for a number of JavaScript applications. For Android smartphone apps and web services written in Java, we propose to develop annotated JML cryptography libraries that can be used with static analysis tools like ESC/Java to verify the security of application code. For clients and web services written in F# for the .NET platform, we propose to use F* to verify their correctness. We also propose to translate verified F* web applications to JavaScript via a verified compiler that preserves the semantics of F* programs in JavaScript.

SECRET Project-Team

3. Research Program

3.1. Scientific foundations

Our research work is mainly devoted to the design and analysis of cryptographic algorithms, either in the classical or in the quantum setting. Our approach on the previous problems relies on a competence whose impact is much wider than cryptology. Our tools come from information theory, discrete mathematics, probabilities, algorithmics, quantum physics... Most of our work mixes fundamental aspects (study of mathematical objects) and practical aspects (cryptanalysis, design of algorithms, implementations). Our research is mainly driven by the belief that discrete mathematics and algorithmics of finite structures form the scientific core of (algorithmic) data protection.

SPADES Team

3. Research Program

3.1. Introduction

The SPADES research program is organized around three main themes, *Components and contracts*, *Real-time multicore programming*, and *Language-based fault tolerance*, that seek to answer the three key questions identified in Section 2.1 . We plan to do so by developing and/or building on programming languages and techniques based on formal methods and formal semantics (hence the use of *"sound programming"* in the project-team title). In particular, we seek to support design where correctness is obtained by construction, relying on proven tools and verified constructs, with programming languages and programming abstractions designed with verification in mind.

3.2. Components and contracts

Component-based construction has long been advocated as a key approach to the "correct-by-construction" design of complex embedded systems [53]. Witness component-based toolsets such as UC Berkeley's Ptolemy [44], Verimag's BIP [30], or the modular architecture frameworks used, for instance, in the automotive industry (AUTOSAR) [25]. For building large, complex systems, a key feature of component-based construction is the ability to associate with components a set of *contracts*, which can be understood as rich behavioral types that can be composed and verified to guarantee a component assemblage will meet desired properties. The goal in this theme is to study the formal foundations of the component-based construction of embedded systems, to develop component and contract theories dealing with real-time, reliability and fault-tolerance aspects of components, and to develop proof-assistant-based tools for the computer-aided design and verification of component-based systems.

Formal models for component-based design are an active area of research (see *e.g.*, [26], [27]). However, we are still missing a comprehensive formal model and its associated behavioral theory able to deal *at the same time* with different forms of composition, dynamic component structures, and quantitative constraints (such as timing, fault-tolerance, or energy consumption). Notions of contracts and interface theories have been proposed to support modular and compositional design of correct-by-construction embedded systems (see *e.g.*, [32], [33] and the references therein), but having a comprehensive theory of contracts that deals with all the above aspects is still an open question [58]. In particular, it is not clear how to accomodate different forms of composition, reliability and fault-tolerance aspects, or to deal with evolving component structures in a theory of contracts.

Dealing in the same component theory with heterogeneous forms of composition, different quantitative aspects, and dynamic configurations, requires to consider together the three elements that comprise a component model: behavior, structure and types. *Behavior* refers to behavioral (interaction and execution) models that characterize the behavior of components and component assemblages (*e.g.*, transition systems and their multiple variants – timed, stochastic, etc.). *Structure* refers to the organization of component assemblages or configurations, and the composition operators they involve. *Types* refer to properties or contracts that can be attached to components and component interfaces to facilitate separate development and ensure the correctness of component configurations with respect to certain properties. Taking into account dynamicity requires to establish an explicit link between behavior and structure, as well as to consider higher-order systems, both of which have a direct impact on types.

We plan to develop our component theory by progressing on two fronts: component calculi, and semantical framework. The work on typed component calculi aims to elicit process calculi that capture the main insights of component-based design and programming and that can serve as a bridge towards actual architecture description and programming language developments. The work on the semantical framework should, in the longer term, provide abstract mathematical models for the more operational and linguistic analysis afforded by component calculi. Our work on component theory will find its application in the development of a Coq-based toolchain for the certified design and construction of dependable embedded systems, which constitutes our third main objective for this axis.

3.3. Real-time multicore programming

Programming real-time systems (i.e. systems whose correct behavior depends on meeting timing constraints) requires appropriate languages (as exemplified by the family of synchronous languages [31]), but also the support of efficient scheduling policies, execution time and schedulability analyses to guarantee real-time constraints (*e.g.*, deadlines) while making the most effective use of available (processing, memory, or networking) resources. Schedulability analysis involves analyzing the worst-case behavior of real-time tasks under a given scheduling algorithm and is crucial to guarantee that time constraints are met in any possible execution of the system. Reactive programming and real-time scheduling and schedulability for multiprocessor systems are old subjects, but they are nowhere as mature as their uniprocessor counterparts, and still feature a number of open research questions [29], [41], in particular in relation with mixed criticality systems. The main goal in this theme is to address several of these open questions.

We intend to focus on two issues: multicriteria scheduling on multiprocessors, and schedulability analysis for real-time multiprocessor systems. Beyond real-time aspects, multiprocessor environments, and multicore ones in particular, are subject to several constraints *in conjunction*, typically involving real-time, reliability and energy-efficiency constraints, making the scheduling problem more complex for both the offline and the online cases. Schedulability analysis for multiprocessor systems, in particular for systems with mixed criticality tasks, is still very much an open research area.

Distributed reactive programming is rightly singled out as a major open issue in the recent, but heavily biased (it essentially ignores recent research in synchronous and dataflow programming), survey by Bainomugisha et al. [29]. For our part, we intend to focus on two questions: devising synchronous programming languages for distributed systems and precision-timed architectures, and devising dataflow languages for multiprocessors supporting dynamicity and parametricity while enjoying effective analyses for meeting real-time, resource and energy constraints in conjunction.

3.4. Language-based fault tolerance

Tolerating faults is a clear and present necessity in networked embedded systems. At the hardware level, modern multicore architectures are manufactured using inherently unreliable technologies [36], [48]. The evolution of embedded systems towards increasingly distributed architectures highlighted in the introductory section means that dealing with partial failures, as in Web-based distributed systems, becomes an important issue. While fault-tolerance is an old and much researched topic, several important questions remain open: automation of fault-tolerance provision, composable abstractions for fault-tolerance, fault diagnosis, and fault isolation.

The first question is related to the old question of “system structure for fault-tolerance” as originally discussed by Randell for software fault tolerance [65], and concerns in part our ability to clearly separate fault-tolerance aspects from the design and programming of purely “functional” aspects of an application. The classical arguments in favor of a clear separation of fault-tolerance concerns from application code revolve around reduced code and maintenance complexity [42]. The second question concerns the definition of appropriate abstractions for the modular construction of fault-tolerant embedded systems. The current set of techniques available for building such systems spans a wide range, including exception handling facilities, transaction management schemes, rollback/recovery schemes, and replication protocols. Unfortunately, these different

techniques do not necessarily compose well – for instance, combining exception handling and transactions is non trivial, witness the flurry of recent work on the topic, see *e.g.*, [52] and the references therein –, they have no common semantical basis, and they suffer from limited programming language support. The third question concerns the identification of causes for faulty behavior in component-based assemblages. It is directly related to the much researched area of fault diagnosis, fault detection and isolation [54].

We intend to address these questions by leveraging programming language techniques (programming constructs, formal semantics, static analyses, program transformations) with the goal to achieve provable fault-tolerance, i.e. the construction of systems whose fault-tolerance can be formally ensured using verification tools and proof assistants. We aim in this axis to address some of the issues raised by the above open questions by using aspect-oriented programming techniques and program transformations to automate the inclusion of fault-tolerance in systems (software as well as hardware), by exploiting reversible programming models to investigate composable recovery abstractions, and by leveraging causality analyses to study fault-ascription in component-based systems. Compared to the huge literature on fault-tolerance in general, in particular in the systems area (see *e.g.*, [49] for an interesting but not so recent survey), we find by comparison much less work exploiting formal language techniques and tools to achieve or support fault-tolerance. The works reported in [34], [37], [39], [46], [55], [64], [69] provide a representative sample of recent such works.

A common theme in this axis is the use and exploitation of causality information. Causality, i.e., the logical dependence of an effect on a cause, has long been studied in disciplines such as philosophy [60], natural sciences, law [61], and statistics [62], but it has only recently emerged as an important focus of research in computer science. The analysis of logical causality has applications in many areas of computer science. For instance, tracking and analyzing logical causality between events in the execution of a concurrent system is required to ensure reversibility [57], to allow the diagnosis of faults in a complex concurrent system [50], or to enforce accountability [56], that is, designing systems in such a way that it can be determined without ambiguity whether a required safety or security property has been violated, and why. More generally, the goal of fault-tolerance can be understood as being to prevent certain causal chains from occurring by designing systems such that each causal chain either has its premises outside of the fault model (*e.g.*, by introducing redundancy [49]), or is broken (*e.g.*, by limiting fault propagation [66]).

SPECFUN Project-Team

3. Research Program

3.1. Studying special functions by computer algebra

Computer algebra manipulates symbolic representations of exact mathematical objects in a computer, in order to perform computations and operations like simplifying expressions and solving equations for "closed-form expressions". The manipulations are often fundamentally of algebraic nature, even when the ultimate goal is analytic. The issue of efficiency is a particular one in computer algebra, owing to the extreme swell of the intermediate values during calculations.

Our view on the domain is that research on the algorithmic manipulation of special functions is anchored between two paradigms:

- adopting linear differential equations as the right data structure for special functions,
- designing efficient algorithms in a complexity-driven way.

It aims at four kinds of algorithmic goals:

- algorithms combining functions,
- functional equations solving,
- multi-precision numerical evaluations,
- guessing heuristics.

This interacts with three domains of research:

- computer algebra, meant as the search for quasi-optimal algorithms for exact algebraic objects,
- symbolic analysis/algebraic analysis;
- experimental mathematics (combinatorics, mathematical physics, ...).

This view is made explicit in the present section.

3.1.1. Equations as a data structure

Numerous special functions satisfy linear differential and/or recurrence equations. Under a mild technical condition, the existence of such equations induces a finiteness property that makes the main properties of the functions decidable. We thus speak of *D-finite functions*. For example, 60 % of the chapters in the handbook [20] describe D-finite functions. In addition, the class is closed under a rich set of algebraic operations. This makes linear functional equations just the right data structure to encode and manipulate special functions. The power of this representation was observed in the early 1990s [72], leading to the design of many algorithms in computer algebra. Both on the theoretical and algorithmic sides, the study of D-finite functions shares much with neighbouring mathematical domains: differential algebra, D-module theory, differential Galois theory, as well as their counterparts for recurrence equations.

3.1.2. Algorithms combining functions

Differential/recurrence equations that define special functions can be recombined [72] to define: additions and products of special functions; compositions of special functions; integrals and sums involving special functions. Zeilberger's fast algorithm for obtaining recurrences satisfied by parametrised binomial sums was developed in the early 1990s already [73]. It is the basis of all modern definite summation and integration algorithms. The theory was made fully rigorous and algorithmic in later works, mostly by a group in RISC (Linz, Austria) and by members of the team [61], [69], [38], [36], [37], [56]. The past ÉPI Algorithms contributed several implementations (*gfun* [64], *Mgfun* [38]).

3.1.3. Solving functional equations

Encoding special functions as defining linear functional equations postpones some of the difficulty of the problems to a delayed solving of equations. But at the same time, solving (for special classes of functions) is a sub-task of many algorithms on special functions, especially so when solving in terms of polynomial or rational functions. A lot of work has been done in this direction in the 1990s; more intensively since the 2000s, solving differential and recurrence equations in terms of special functions has also been investigated.

3.1.4. Multi-precision numerical evaluation

A major conceptual and algorithmic difference exists for numerical calculations between data structures that fit on a machine word and data structures of arbitrary length, that is, *multi-precision* arithmetic. When multi-precision floating-point numbers became available, early works on the evaluation of special functions were just promising that "most" digits in the output were correct, and performed by heuristically increasing precision during intermediate calculations, without intended rigour. The original theory has evolved in a twofold way since the 1990s: by making computable all constants hidden in asymptotic approximations, it became possible to guarantee a *prescribed* absolute precision; by employing state-of-the-art algorithms on polynomials, matrices, etc, it became possible to have evaluation algorithms in a time complexity that is linear in the output size, with a constant that is not more than a few units. On the implementation side, several original works exist, one of which (*NumGfun* [60]) is used in our DDMF.

3.1.5. Guessing heuristics

"Differential approximation", or "Guessing", is an operation to get an ODE likely to be satisfied by a given approximate series expansion of an unknown function. This has been used at least since the 1970s and is a key stone in spectacular applications in experimental mathematics [34]. All this is based on subtle algorithms for Hermite–Padé approximants [24]. Moreover, guessing can at times be complemented by proven quantitative results that turn the heuristics into an algorithm [32]. This is a promising algorithmic approach that deserves more attention than it has received so far.

3.1.6. Complexity-driven design of algorithms

The main concern of computer algebra has long been to prove the feasibility of a given problem, that is, to show the existence of an algorithmic solution for it. However, with the advent of faster and faster computers, complexity results have ceased to be of theoretical interest only. Nowadays, a large track of works in computer algebra is interested in developing fast algorithms, with time complexity as close as possible to linear in their output size. After most of the more pervasive objects like integers, polynomials, and matrices have been endowed with fast algorithms for the main operations on them [43], the community, including ourselves, started to turn its attention to differential and recurrence objects in the 2000s. The subject is still not as developed as in the commutative case, and a major challenge remains to understand the combinatorics behind summation and integration. On the methodological side, several paradigms occur repeatedly in fast algorithms: "divide and conquer" to balance calculations, "evaluation and interpolation" to avoid intermediate swell of data, etc. [29].

3.2. Trusted computer-algebra calculations

3.2.1. Encyclopedias

Handbooks collecting mathematical properties aim at serving as reference, therefore trusted, documents. The decision of several authors or maintainers of such knowledge bases to move from paper books [20], [22], [65] to websites and wikis [0] allows for a more collaborative effort in proof reading. Another step toward further confidence is to manage to generate the content of an encyclopedia by computer-algebra programs, as is the case with the Wolfram Functions Site [0] or DDMF [0]. Yet, due to the lingering doubts about computer-algebra systems, some encyclopedias propose both cross-checking by different systems and handwritten companion paper proofs of their content [0]. As of today, there is no encyclopedia certified with formal proofs.

[0] for instance http://dlmf.nist.gov/ for special functions or http://oeis.org/ for integer sequences
[0] http://functions.wolfram.com/
[0] http://ddmf.msr-inria.inria.fr/1.9.1/ddmf

3.2.2. Computer algebra and symbolic logic

Several attempts have been made in order to extend existing computer-algebra systems with symbolic manipulations of logical formulas. Yet, these works are more about extending the expressivity of computer-algebra systems than about improving the standards of correctness and semantics of the systems. Conversely, several projects have addressed the communication of a proof system with a computer-algebra system, resulting in an increased automation available in the proof system, to the price of the uncertainty of the computations performed by this oracle.

3.2.3. Certifying systems for computer algebra

More ambitious projects have tried to design a new computer-algebra system providing an environment where the user could both program efficiently and elaborate formal and machine-checked proofs of correctness, by calling a general-purpose proof assistant like the Coq system. This approach requires a huge manpower and a daunting effort in order to re-implement a complete computer-algebra system, as well as the libraries of formal mathematics required by such formal proofs.

3.2.4. Semantics for computer algebra

The move to machine-checked proofs of the mathematical correctness of the output of computer-algebra implementations demands a prior clarification about the often implicit assumptions on which the presumably correctly implemented algorithms rely. Interestingly, this preliminary work, which could be considered as independent from a formal certification project, is seldom precise or even available in the literature.

3.2.5. Formal proofs for symbolic components of computer-algebra systems

A number of authors have investigated ways to organize the communication of a chosen computer-algebra system with a chosen proof assistant in order to certify specific components of the computer-algebra systems, experimenting various combinations of systems and various formats for mathematical exchanges. Another line of research consists in the implementation and certification of computer-algebra algorithms inside the logic [68], [48], [57] or as a proof-automation strategy. Normalization algorithms are of special interest when they allow to check results possibly obtained by an external computer-algebra oracle [41]. A discussion about the systematic separation of the search for a solution and the checking of the solution is already clearly outlined in [54].

3.2.6. Formal proofs for numerical components of computer-algebra systems

Significant progress has been made in the certification of numerical applications by formal proofs. Libraries formalizing and implementing floating-point arithmetic as well as large numbers and arbitrary-precision arithmetic are available. These libraries are used to certify floating-point programs, implementations of mathematical functions and for applications like hybrid systems.

3.3. Machine-checked proofs of formalized mathematics

To be checked by a machine, a proof needs to be expressed in a constrained, relatively simple formal language. Proof assistants provide facilities to write proofs in such languages. But, as merely writing, even in a formal language, does not constitute a formal proof just per se, proof assistants also provide a proof checker: a small and well-understood piece of software in charge of verifying the correctness of arbitrarily large proofs. The gap between the low-level formal language a machine can check and the sophistication of an average page of mathematics is conspicuous and unavoidable. Proof assistants try to bridge this gap by offering facilities, like notations or automation, to support convenient formalization methodologies. Indeed, many aspects, from the logical foundation to the user interface, play an important role in the feasibility of formalized mathematics inside a proof assistant.

[0] http://129.81.170.14/~vhm/Table.html

3.3.1. Logical foundations and proof assistants

While many logical foundations for mathematics have been proposed, studied, and implemented, type theory is the one that has been more successfully employed to formalize mathematics, to the notable exception of the Mizar system [58], which is based on set theory. In particular, the calculus of construction (CoC) [39] and its extension with inductive types (CIC) [40], have been studied for more than 20 years and been implemented by several independent tools (like Lego, Matita, and Agda). Its reference implementation, Coq [66], has been used for several large-scale formalizations projects (formal certification of a compiler back-end; four-color theorem). Improving the type theory underlying the Coq system remains an active area of research. Other systems based on different type theories do exist and, whilst being more oriented toward software verification, have been also used to verify results of mainstream mathematics (prime-number theorem; Kepler conjecture).

3.3.2. Computations in formal proofs

The most distinguishing feature of CoC is that computation is promoted to the status of rigorous logical argument. Moreover, in its extension CIC, we can recognize the key ingredients of a functional programming language like inductive types, pattern matching, and recursive functions. Indeed, one can program effectively inside tools based on CIC like Coq. This possibility has paved the way to many effective formalization techniques that were essential to the most impressive formalizations made in CIC.

Another milestone in the promotion of the computations-as-proofs feature of Coq has been the integration of compilation techniques in the system to speed up evaluation. Coq can now run realistic programs in the logic, and hence easily incorporates calculations into proofs that demand heavy computational steps.

Because of their different choice for the underlying logic, other proof assistants have to simulate computations outside the formal system, and indeed fewer attempts to formalize mathematical proofs involving heavy calculations have been made in these tools. The only notable exception, which was finished in 2014, the Kepler conjecture, required a significant work to optimize the rewriting engine that simulates evaluation in Isabelle/HOL.

3.3.3. Large-scale computations for proofs inside the Coq system

Programs run and proved correct inside the logic are especially useful for the conception of automated decision procedures. To this end, inductive types are used as an internal language for the description of mathematical objects by their syntax, thus enabling programs to reason and compute by case analysis and recursion on symbolic expressions.

The output of complex and optimized programs external to the proof assistant can also be stamped with a formal proof of correctness when their result is easier to *check* than to *find*. In that case one can benefit from their efficiency without compromising the level of confidence on their output at the price of writing and certify a checker inside the logic. This approach, which has been successfully used in various contexts, is very relevant to the present research project.

3.3.4. Relevant contributions from the Mathematical Component libraries

Representing abstract algebra in a proof assistant has been studied for long. The libraries developed by the MathComp project for the proof of the Odd Order Theorem provide a rather comprehensive hierarchy of structures; however, they originally feature a large number of instances of structures that they need to organize. On the methodological side, this hierarchy is an incarnation of an original work [42] based on various mechanisms, primarily type inference, typically employed in the area of programming languages. A large amount of information that is implicit in handwritten proofs, and that must become explicit at formalization time, can be systematically recovered following this methodology.

Small-scale reflection [45] is another methodology promoted by the MathComp project. Its ultimate goal is to ease formal proofs by systematically dealing with as many bureaucratic steps as possible, by automated computation. For instance, as opposed to the style advocated by Coq's standard library, decidable predicates are systematically represented using computable boolean functions: comparison on integers is expressed as

program, and to state that $a \leq b$ one compares the output of this program run on a and b with $true$. In many cases, for example when a and b are values, one can prove or disprove the inequality by pure computation.

The MathComp library was consistently designed after uniform principles of software engineering. These principles range from simple ones, like naming conventions, to more advanced ones, like generic programming, resulting in a robust and reusable collection of formal mathematical components. This large body of formalized mathematics covers a broad panel of algebraic theories, including of course advanced topics of finite group theory, but also linear algebra, commutative algebra, Galois theory, and representation theory. We refer the interested reader to the online documentation of these libraries [67], which represent about 150,000 lines of code and include roughly 4,000 definitions and 13,000 theorems.

Topics not addressed by these libraries and that might be relevant to the present project include real analysis and differential equations. The most advanced work of formalization on these domains is available in the HOL-Light system [50], [51], [52], although some existing developments of interest [27], [59] are also available for Coq. Another aspect of the MathComp libraries that needs improvement, owing to the size of the data we manipulate, is the connection with efficient data structures and implementations, which only starts to be explored.

3.3.5. User interaction with the proof assistant

The user of a proof assistant describes the proof he wants to formalize in the system using a textual language. Depending on the peculiarities of the formal system and the applicative domain, different proof languages have been developed. Some proof assistants promote the use of a declarative language, when the Coq and Matita systems are more oriented toward a procedural style.

The development of the large, consistent body of MathComp libraries has prompted the need to design an alternative and coherent language extension for the Coq proof assistant [47], [46], enforcing the robustness of proof scripts to the numerous changes induced by code refactoring and enhancing the support for the methodology of small-scale reflection.

The development of large libraries is quite a novelty for the Coq system. In particular any long-term development process requires the iteration of many refactoring steps and very little support is provided by most proof assistants, with the notable exception of Mizar [63]. For the Coq system, this is an active area of research.

SUMO Project-Team

3. Research Program

3.1. Model expressivity and quantitative verification

The overall objective of this axis is to combine the quantitative aspects of models with a distributed/modular setting, while maintaining the tractability of verification and management objectives.

There is first an issue of modeling, to nicely weave time, costs and probabilities with concurrency and/or asynchronism. Several approaches are quite natural, as time(d) Petri nets, networks of timed automata, communicating synchronously or through FIFO, etc. But numerous bottlenecks remain. For example, so far, no probabilistic model nicely fits the notion of concurrency: there is no clean way to express that two components are stochastically independent between two rendez-vous.

Second, the models we want to manipulate should allow for quantitative verification. This covers two aspects: either the verification question is itself quantitative (compute an optimal scheduling policy) or boolean (decide whether the probability is greater than a threshold). Our goal is to explore the frontier between decidable and undecidable problems, or more pragmatically tractable and untractable problems. Of course, there is a tradeoff between the expressivity and the tractability of a model. Models that incorporate distributed aspects, probabilities, time, etc, are typically untractable. In such a case, abstraction or approximation techniques are a work around that we will explore.

In more details, our research program on this axis covers the following topics:

- the verification of distributed timed systems,
- the verification of large scale probabilistic (dynamic) systems, with a focus on approximation techniques for such systems,
- the evaluation of the opacity/diagnosability degree of stochastic systems,
- the design of modular testing methods for large scale modular systems.

3.2. Management of large distributed systems

The generic terms of "supervision" or "management" of distributed systems cover problems like control (and controller synthesis), diagnosis, sensor placement, planning, optimization, (state) estimation, parameter identification, testing, etc. These questions have both an offline and an online facet. The literature is abundant for discrete event systems (DES), even in the distributed case, and for some quantitative aspects of DES in the centralized case (for example partially observed Markov decision processes (POMDP), probabilistic diagnosis/diagnosers, (max,+) approaches to timed automata). And there is a strong trend driving formal methods approaches towards quantitative models and questions like the most likely diagnosis, control for best average reward or for best QoS, optimal sensor placement, computing the probability of failure (un)detection, estimating the average impact of some failure or of a decision, etc. This second research axis focuses on these issues, and aims at developing new concepts and tools to master some already existing large scale systems, as telecommunication networks, cloud infrastructures, web-services, etc. (see the Application Domains section).

The objective being to address large systems, our work will be driven by two considerations: how to take advantage of the modularity of systems, and how to best approximate/abstract too complex systems by more tractable ones. We mention below main topics we will focus on:

- Approximate management methods. We will explore the relevance of ideas developed for large scale stochastic systems, as turbo-algorithms for example, in the setting of modular dynamic systems.
- Self-modeling, which consists in managing large scale systems that are known by their building rules, but which specific managed instance is only discovered at runtime, and on the fly. The model of the managed system is built on-line, following the needs of the management algorithms.

- Distributed control. We will tackle issues related to asynchronous communications between local controllers, and abstraction techniques to address large systems.
- Test and enforcement. We will tackle coverage issues for the test of large systems, and the test and enforcement of properties for timed models, or for systems handling data.

3.3. Data driven systems

The term data-driven systems refers to systems the behavior of which depends both on explicit workflows (scheduling and durations of tasks, calls to possibly distant services,...) and on the data processed by the system (stored data, parameters of a request, results of a request,...). This family of systems covers workflows that convey data (business processes or information systems), transactional systems (web stores), large databases managed with rules (banking systems), collaborative environments (health systems), etc. These systems are distributed, modular, and open: they integrate components and sub-services distributed over the web and accept requests from clients. Our objective is to provide validation and supervision tools for such systems. To achieve this goal, we have to solve several challenging tasks:

- provide realistic models, and sound automated abstraction techniques, to reason on models that are reasonable abstractions of real implemented systems designed in low-level languages (for instance BPEL (Business Process Execution Language)). These models should be able to encompass modularity, distribution, in a context where workflows and data aspects are tightly connected.
- provide tractable solutions for validation of models. Important questions that are frequently addressed (for instance safety properties or coverability) should remain decidable on our models, but also with a decent complexity.
- address design of data driven systems in a declarative way: declarative models are another way to handle data-driven systems. Rather than defining the explicit workflows and their effects on data, rule-based models state how actions are enacted in terms of the shape (pattern matching) or value of the current data. Such declarative models are well accepted in business processes (Companies such as IBM use their own model of business rules [53] to interact with their clients). Our approach is to design collaborative activities in terms of distributed structured documents, that can be seen as communicating rewriting systems. This modeling paradigm also includes models such as distributed Active XML [48], [51]. We think that distributed rewriting rules or attributed grammars can provide a practical but yet formal framework for maintenance, by providing a solution to update mandatory documentation during the lifetime of an artifact.
- address QoS management in large reconfigurable systems:

Data driven distributed systems such as web services often have constraints in terms of QoS. This calls for an analysis of quantitative features, and for reconfiguration techniques to meet QoS contracts. We will build from the experience in our team on QoS contracts composition [54] and planning [47], [49] to propose optimization and reconfiguration schemes.

TASC Project-Team

3. Research Program

3.1. Overview

Basic research is guided by the challenges raised before: to classify and enrich the models, to automate reformulation and resolution, to dissociate declarative and procedural knowledge, to come up with theories and tools that can handle problems involving both continuous and discrete variables, to develop modelling tools and to come up with solving tools that scale well. On the one hand, **classification aspects** of this research are integrated within a knowledge base about combinatorial problem solving: the global constraint catalog (see http://sofdem.github.io/gccat/). On the other hand, **solving aspects** are capitalized within the constraint solving system CHOCO. Lastly, within the framework of its activities of valorisation, teaching and of partnership research, the team uses constraint programming for solving various concrete problems. The challenge is, on one side to increase the visibility of the constraints in the others disciplines of computer science, and on the other side to contribute to a broader diffusion of the constraint programming in the industry.

3.2. Fundamental Research Topics

This part presents the research topics investigated by the project:

- Global Constraints Classification, Reformulation and Filtering,
- Convergence between Discrete and Continuous,
- Dynamic, Interactive and over Constrained Problems,
- Solvers.

These research topics are in fact not independent. The work of the team thus frequently relates transverse aspects such as explained global constraints, Benders decomposition and explanations, flexible and dynamic constraints, linear models and relaxations of constraints.

3.2.1. Constraints Classification, Reformulation and Filtering

In this context our research is focused (a) first on identifying recurring combinatorial structures that can be used for modelling a large variety of optimization problems, and (b) exploit these combinatorial structures in order to come up with efficient algorithms in the different fields of optimization technology. The key idea for achieving point (b) is that many filtering algorithms both in the context of Constraint Programming, Mathematical Programming and Local Search can be interpreted as the maintenance of invariants on specific domains (e.g., graph, geometry). The systematic classification of global constraints and of their relaxation brings a synthetic view of the field. It establishes links between the properties of the concepts used to describe constraints and the properties of the constraints themselves. Together with SICS, the team develops and maintains *a catalog of global constraints*, which describes the semantics of more than 431 constraints, and proposes a unified mathematical model for expressing them. This model is based on graphs, automata and logic formulae and allows to derive filtering methods and automatic reformulation for each constraint in a unified way (see http://www.emn.fr/x-info/sdemasse/gccat/index.html). We consider hybrid methods (i.e., methods that involve more than one optimization technology such as constraint programming, mathematical programming or local search), to draw benefit from the respective advantages of the combined approaches. More fundamentally, the study of hybrid methods makes it possible to compare and connect strategies of resolution specific to each approach for then conceiving new strategies. Beside the works on classical, complete resolution techniques, we also investigate local search techniques from a mathematical point of view. These partly random algorithms have been proven very efficient in practice, although we have little theoretical knowledge on their behaviour, which often makes them problem-specific. Our research in that area is focused on a probabilistic model of local search techniques, from which we want to derive quantified information on their behaviour, in order to

use this information directly when designing the algorithms and exploit their performances better. We also consider algorithms that maintain local and global consistencies, for more specific models. Having in mind the trade off between genericity and effectiveness, the effort is put on the efficiency of the algorithms with guarantee on the produced levels of filtering. This effort results in adapting existing techniques of resolution such as graph algorithms. For this purpose we identify necessary conditions of feasibility that can be evaluated by efficient incremental algorithms. Genericity is not neglected in these approaches: on the one hand the constraints we focus on are applicable in many contexts (for example, graph partitioning constraints can be used both in logistics and in phylogeny); on the other hand, this work led to study the portability of such constraints and their independence with specific solvers. This research orientation gathers various work such as strong local consistencies, graph partitioning constraints, geometrical constraints, and optimization and soft constraints. Within the perspective to deal with complex industrial problems, we currently develop meta constraints (e.g. *geost*) handling all together the issues of large-scale problems, dynamic constraints, combination of spatial and temporal dimensions, expression of business rules described with first order logic.

3.2.2. Convergence between Discrete and Continuous

Many industrial problems mix continuous and discrete aspects that respectively correspond to physical (e.g., the position, the speed of an object) and logical (e.g., the identifier, the nature of an object) elements. Typical examples of problems are for instance:

- *Geometrical placement problems* where one has to place in space a set of objects subject to various geometrical constraints (i.e., non-overlapping, distance). In this context, even if the positions of the objects are continuous, the structure of optimal configurations has a discrete nature.
- *Trajectory and mission planning problems* where one has to plan and synchronize the moves of several teams in order to achieve some common goal (i.e., fire fighting, coordination of search in the context of rescue missions, surveillance missions of restricted or large areas).
- *Localization problems in mobile robotic* where a robot has to plan alone (only with its own sensors) its trajectory. This kind of problematic occurs in situations where the GPS cannot be used (e.g., under water or Mars exploration) or when it is not precise enough (e.g., indoor surveillance, observation of contaminated sites).

Beside numerical constraints that mix continuous and integer variables we also have global constraints that involve both type of variables. They typically correspond to graph problems (i.e., graph colouring, domination in a graph) where a graph is dynamically constructed with respect to geometrical and-or temporal constraints. In this context, the key challenge is avoiding decomposing the problem in a discrete and continuous parts as it is traditionally the case. As an illustrative example consider *the wireless network deployment problem*. On the one hand, the continuous part consists of finding out where to place a set of antenna subject to various geometrical constraints. On the other hand, by building an interference graph from the positions of the antenna, the discrete part consists of allocating frequencies to antenna in order to avoid interference. In the context of convergence between discrete and continuous variables, our goals are:

- First to identify and compare typical class of techniques that are used in the context of continuous and discrete solvers.
- To see how one can unify and/or generalize these techniques in order to handle in an integrated way continuous and discrete constraints within the same framework.

3.2.3. Dynamic, Interactive and over Constrained Problems

Some industrial applications are defined by a set of constraints which may change over time, for instance due to an interaction with the user. Many other industrial applications are over-constrained, that is, they are defined by set of constraints which are more or less important and cannot be all satisfied at the same time. Generic, dedicated and explanation-based techniques can be used to deal efficiently with such applications. Especially, these applications rely on the notion of *soft constraints* that are allowed to be (partially) violated. The generic concept that captures a wide variety of soft constraints is the violation measure, which is coupled with specific resolution techniques. Lastly, soft constraints allow to combine the expressive power of global constraints with local search frameworks.

3.2.4. Solvers

- *Discrete solver* Our theoretical work is systematically validated by concrete experimentations. We have in particular for that purpose the CHOCO constraint platform. The team develops and maintains CHOCO initially with the assistance of the laboratory e-lab of Bouygues (G. Rochart), the company Amadeus (F. Laburthe), and others researchers such as H. Cambazard (4C, INP Grenoble). Since 2008 the main developments are done by Charles Prud'homme and Xavier Lorca. The functionalities of CHOCO are gradually extended with the outcomes of our works: design of constraints, analysis and visualization of explanations, etc. The open source CHOCO library is downloaded on average 450 times each month since 2006. CHOCO is developed in line with the research direction of the team, in an open-minded scientific spirit. Contrarily to other solvers where the efficiency often relies on problem-specific algorithms, CHOCO aims at providing the users both with reusable techniques (based on an up-to-date implementation of the global constraint catalogue) and with a variety of tools to ease the use of these techniques (clear separation between model and resolution, event-based solver, management of the over-constrained problems, explanations, etc.).
- *Continuous solver* Since 2009 year, due to the hiring of Gilles Chabert, the team is also involved in the development of the continuous constraint solver IBEX. These developments led us to new research topics, suitable for the implementation of discrete and continuous constraint solving systems: portability of the constraints, management of explanations, incrementality and recalculation. They partially use aspect programming (in collaboration with the InriaASCOLA team).
- *Constraint programming and verification* Constraint Programming has already had several applications to verification problems. It also has many common ideas with Abstract Interpretation, a theory of approximation of the semantics of programs. In both cases, we are interested in a particular set (solutions in CP, program traces in semantics), which is hard or impossible to compute, and this set is replaced by an over-approximation (consistent domains / abstract domains). Previous works (internship of Julie Laniau, PhD of Marie Pelleau, collaboration with the Abstract Interpretation team at the ENS and Antoine Miné in particular) have exhibited some of these links, and identified some situations where the two fields, Abstract Interpretation and Constraint Programming, can complement each other. It is the case in real-time stream processing languages, where Abstract Interpretation techniques may not be precise enough when analyzing loops. With the PhD of Anicet Bart, we are currently working on using CP techniques to find loop invariants for the Faust language, a functional sound processing language.

This work around the design and the development of solvers thus forms the fourth direction of basic research of the project.

TEA Project-Team

3. Research Program

3.1. State of the Art

System design based on the “synchronous paradigm” has focused the attention of many academic and industrial actors on abstracting non-functional implementation details from system design. This design abstraction focuses on the logic of interaction in reactive programs rather than their timed behaviour, allowing to secure functional correctness while remaining an intuitive programming model for embedded systems.

Maintaining the “synchronous hypothesis” on software at runtime, however, demands a quasi-synchronous model of execution (hardware or middleware) in order to be effectively implemented [0]. Strong software constraints to ensure functional correctness imply strong runtime restrictions and simple hardware. If we look at recent features found in synchronous programming languages such as Quartz [0], Lucid [0] departing from the simpler semantics of Esterel [0] and Lustre [0], we observe that all try to cope in a way or another with the availability of more general execution architectures: clock domains [0], pipelining [0], streaming [0]. Unfortunately, attempts to scale the simple "typed programming language" approach of the 90’s [0] to the above purpose hit inherent computational complexity limits. For example, a periodic clock operation like $0^{(1920*(1080-480))}\{0^{1200}1^{720}\}^{480}$ in Lucy-n (0^n means n zeros) yields an exponentially larger term [0]. This explains why team TEA opts for focusing on the semantics of time and concurrency in system design and on implementing the implied design methodologies using program analysis and abstract interpretation.

By contrast with a synchronous hypothesis, the polychronous MoCC implemented in the specification language Signal, available in the Eclipse project POP [0] and in the CCSL standard [0], is inherently capable of describing circuits and systems with multiple clocks.

The Eclipse project POP provides a tooled infrastructure to refine high-level specifications into real-time streaming applications or locally synchronous and globally asynchronous systems, through a series of model analysis and synthesis libraries. These tool-supported refinement and transformation techniques can assist the system engineer from the earliest design stages of requirement specification to the latest stages of synthesis, scheduling and deployment. These characteristics make polychrony much closer to the required semantic for compositional, refinement-based, architecture-driven, system design.

3.2. Modelling Time

The elegant abstraction offered by the "synchronous hypothesis" [0] has translated in famous leitmotivs like *"computation takes no time"* and *"communication is instantaneous"* and contributed to the impact and commercial success of Esterel Studio [0] and SCADE [0].

[0] *A protocol for loosely time-triggered architectures*. A. Benveniste et al. Embedded Software Conference. ACM, 2002
[0] The Averest System http://www.averest.org.
[0] Lucid synchrone http://www.di.ens.fr/~pouzet/lucid-synchrone.
[0] *The Esterel synchronous programming language*. G. Berry, G. Gonthier. Science of Computer Programming, v. 19(2). Elsevier, 1992.
[0] *The synchronous data flow programming language Lustre*. Halbwachs, N., Caspi, P., Raymond, P., Pilaud, D. Proceedings of the IEEE v. 79(9), 1991.
[0] *A formal semantics of clock refinement in imperative synchronous languages*. Gemünde, M., Brandt, J., Schneider, K. Application of Concurrency to System Design. IEEE Press, 2010.
[0] *Parallelism with futures in Lustre*. Cohen, A., Gérard, L., Pouzet, M. Embedded Software Conference. ACM, 2012.
[0] *N-synchronous Kahn networks*. Cohen, A., et al. Principles of Programming Languages. ACM, 2006.
[0] A. Benveniste et al. *The Synchronous Languages Twelve Years Later*. Proceedings of the IEEE v. 91(1), 2003.
[0] http://www.di.ens.fr/~guatto/slides_parkas_14_05_12.pdf, page 15.
[0] *Polychrony on POLARSYS (POP)*, an Eclipse project in the POLARSYS Industry Working Group, 2013. https://www.POLARSYS.org/projects/POLARSYS.pop
[0] *Clock Constraints in UML/MARTE CCSL*. C. André, F. Mallet. Technical Report RR-6540. Inria, 2008. http://hal.inria.fr/inria-00280941
[0] *The synchronous languages 12 years later*. A. Benveniste, et al. Proceedings of IEEE, 91(1), 2003.
[0] *Esterel Studio*, Sinfora. http://www.synfora.com/products/esterelStudio.html.

Meanwhile, proposals and standards have appeared to push the technical boundaries of synchronous concurrency, in order to address a larger spectrum of concerns related to modern, heterogeneous, many-core architectures. The challenge becomes more largely about representing time in system design, alongside with many, so called, non-functional properties: cost, power, heat, speed, throughput.

One reference for the purpose of modelling timed hardware behaviour is PSL [0]. PSL is a formal specification language based on Kleene algebras that was originally designed to model regular hardware signal traces. The duality between automata and this formalism also makes it suitable to express requirements, formal properties and abstraction of program behaviours. It is widely used for modelling and verification of hardware systems.

A more recent reference of broader spectrum is CCSL [0], the clock constraints specification language of UML Marte. CCSL's core specification formalism is based on the Signal MoCC, it is synchronous and multi-clocked. Yet, CCSL supports extensions to model multi-rate, multi-periodic systems, that are adequate to represent hardware clocks, as well as asynchronous and continuous extensions (although largely unexploited in the related work). Another well-developed model is that of Ptolemy [0], which represents time as a first-class citizen alongside data carried by streams in the modelled system. It relates to the notion of PRET [0], (precision time machine) to support real-time simulation.

In the meantime, and from a totally different perspective, type theory has made considerable advances since the avent of effect systems [0] to formally represent formal properties alongside with values. Hybrid types [0] (linked to interface and contract theories), refinement types [0], value-dependant types, allow formal program properties, logical or temporal, to flow alongside with data-types during program analysis and verification. While a combination of all the above is yet unexplored, it offers an exciting venue for contributing in either/both of these fields with new theoretical developments on modelling time using principles of type theory.

3.3. Modelling Architectures

An architectural model represents components in a distributed system as boxes with well-defined interfaces, connections between ports on component interfaces, and specifies component properties that can be used in analytical reasoning about the model. Models are hierarchically organised, so that each box can contain another system with its own set of boxes and connections between them. An architecture description language for embedded systems, for which timing and resource availability form an important part of the requirements, must in addition describe resources of the system platform, such as processors, memories, communication links, etc. Several architectural modelling languages for embedded systems have emerged in recent years, including the SAE AADL [0], SysML [0], UML MARTE [0].

An architectural specification serves several important purposes. First, it breaks down a system model into manageable components to establish clear interfaces between components. In this way, complexity becomes manageable by hiding details that are not relevant at a given level of abstraction. Clear, formally defined, component interfaces allow us to avoid integration problems at the implementation phase. Connections between components, which specify how components affect each other, help propagate the effects of a change in one component to the linked components.

[0] *Scade System*, ANSYS. http://www.esterel-technologies.com/products/scade-system
[0] *IEEE Standard for Property Specification Language*. IEEE, 2005. http://dx.doi.org/10.1109/IEEESTD.2005.97780.
[0] *CCSL: specifying clock constraints with UML/MARTE*, OMG, 2008. http://www.omgmarte.org/node/66.
[0] *Ptolemy*, UC Berkeley. http://ptolemy.eecs.berkeley.edu.
[0] *Precision Timed Computation in Cyber-Physical Systems*. E. A. Lee and S. A. Edwards, 2007. http://ptolemy.eecs.berkeley.edu/publications/papers/07/PRET.
[0] *Polymorphic effect systems*. J. M. Lucassen , D. K. Gifford. Principles of Programming Languages. ACM, 1988.
[0] *Hybrid type checking*. K.W. Knowles and C. Flanagan. ACM Transactions on Programming languages and systems, 32(2). ACM, 2010
[0] *Abstract Refinement Types*. N. Vazou, P. Rondon, and R. Jhala. European Symposium on Programming. Springer, 2013.
[0] *Architecture Analysis and Design Language*, AS-5506. SAE, 2004. http://standards.sae.org/as5506b
[0] *System Modelling Language*. OMG, 2007. http://www.omg.org/spec/SysML
[0] *UML Profile for MARTE*. OMG, 2009. http://www.omg.org/spec/MARTE

Most importantly, an architectural model is a repository to share knowledge about the system being designed. This knowledge can be represented as requirements, design artefacts, component implementations, held together by a structural backbone. Such a repository enables automatic generation of analytical models for different aspects of the system, such as timing, reliability, security, performance, energy, etc. Since all the models are generated from the same source, the consistency of assumptions w.r.t. guarantees, of abstractions w.r.t. refinements, used for different analyses becomes easier, and can be properly ensured in a design methodology based on formal verification and synthesis methods.

Related works in this aim, and closer in spirit to our approach (to focus on modelling time) are domain-specific languages such as Prelude [0] to model the real-time characteristics of embedded software architectures. Conversely, standard architecture description languages could be based on algebraic modelling tools, such as interface theories with the ECDAR tool [0].

3.4. Time Scheduling

Cyber-physical systems are reactive systems whose correctness not only depends on a deterministic behavior but also on timing predictability. The timing parameters of a CPS are requirements that arise from the system's specification (e.g. minimum throughput, maximum latency, deadlines) or timing properties of the physical and cyber-parts that restrict the CPS implementation. The design of a CPS must ensure that these timing requirements will be met, even in the worst-case scenario, through the different components and their timing properties.

3.4.1. Scheduling theory

Real-time scheduling theory provides tools for predicting the timing behaviour of a CPS which consists of many interacting software and hardware components. Expressing parallelism among software components is a crucial aspect of the design process of a CPS. It allows for efficient partition and exploitation of available resources. In the real-time scheduling theory literature, many models of computation have been proposed to express such parallelism, for instance:

- Set of independent periodic, sporadic, or aperiodic tasks where each real-time task is generally characterised with some timing parameters: deadline, period, first start time, jitter, etc. The periodic and sporadic task models [0] are very well studied task models since they allow to analytically reason about the timing behaviour of tasks. More expressive task models [0] such as the multi-frame and the recurring real-time task models have also emerged.
- Task graph models [0] where precedence constraints among real-time tasks may exist.
- Data-flow graph models such as synchronous data-flow (SDF [0]) and cyclo-static dataflow (CSDF [0]. IEEE, 1996.) models where the set of tasks (also called actors) communicate with each other through FIFO channels. When it fires, an actor consumes a predefined number of tokens from its inputs and produces a predefined number of tokens on its outputs. The scheduling problem is hence more complex since data dependencies must be satisfied.

[0] *The Prelude language*. LIFL and ONERA, 2012. http://www.lifl.fr/~forget/prelude.html
[0] *PyECDAR, timed games for timed specifications*. Inria, 2013. https://project.inria.fr/pyecdar
[0] *Scheduling algorithms for multiprogramming in a hard-real-time environment*. C. L. Liu and J. W. Layland. Journal of the ACM 20(1), 1973.
[0] *The digraph real-time task model*. M. Stigge, P. Ekberg, N. Guan, and W. Yi. Real-Time and Embedded Technology and Applications Symposium. IEEE, 2011.
[0] *Task graph scheduling using timed automata*. Y. Abdeddaïm, A. Kerbaa, and O. Maler. International Symposium on Parallel and Distributed Processing. IEEE, 2003.
[0] *Synchronous data-flow*. E. A. Lee and D. G. Messerschmitt. Proceedings of the IEEE, 1987.
[0] *Cycle-static dataflow*. G. Blisen, M. Engels, R. Lauwereins, and J. Peperstraete. Transactions on Signal Processing

The literature about real-time scheduling of sets of independent real-time tasks [0] provides very mature schedulability tests regarding many scheduling strategies, preemptive or non-preemptive scheduling, uniprocessor or multiprocessor scheduling, etc. Historically, real-time systems where scheduled by cyclic executives (i.e. static scheduling). However, since this approach produces rigid and difficult to maintain systems and handles only periodic tasks, the research community has proposed many dynamic scheduling algorithms, which can be classified as fixed-priority scheduling (e.g. rate-monotonic scheduling, deadline monotonic scheduling) and dynamic priority scheduling (e.g. earliest-deadline first scheduling, least laxity scheduling). Multiprocessor scheduling can be further classified as partitioned scheduling (each task is allocated to a processor and no migration is allowed), global scheduling (a single job can migrate to and execute on different processors), or hybrid.

Scheduling of data-flow graphs has also been extensively studied in the past decades. Static-periodic scheduling is the main scheduling approach, which consists in infinitely repeating a firing sequence of actors. This problem has been addressed with respect to many performance criteria: throughput maximisation [0], latency minimisation [0], buffer minimisation [0], code size minimisation [0], etc. Recently, real-time dynamic scheduling (fixed-priority and earliest-deadline first scheduling) of data-flow graphs has been addressed where actors are mapped to periodic real-time tasks and existing schedulability tests are adapted to synthesise the timing characteristics of actors [00]

3.5. Virtual Prototyping

Virtual Prototyping is the technology of developing realistic simulators from models of a system under design; that is, an emulated device that captures most, if not all, of the required properties of the real system, based on its specifications. A virtual prototype should be run and tested like the real device. Ideally, the real application software would be run on the virtual prototyping platform and produce the same results as the real device with the same sequence of outputs and reported performance measurements. This may be true to some extent only. Some trade-offs have often to be made between the accuracy of the virtual prototype, and time to develop accurate models.

A virtual prototyping platform must include operating system or hardware emulation technology since the hardware functions must be simulated at least to a minimum extent in order to run the software and evaluate the design alternatives. The hardware simulation engine is a key component of a virtual prototyping platform, which makes it possible to run the application software and produce output that can be analysed by other tools. Because electronic design tools (EDAs) simulate the hardware in every detail, it is possible to verify that the circuit operates properly and also to measure how many clock cycles will be required to achieve an operation. But because they simulate very low-level operations, simulation is much too slow to be usable for virtual prototyping. The authors of the FAST system [0] and SocLib project reports [0] speed-ups with a factor of several

[0] *A survey of hard real-time scheduling for multiprocessor systems*. R. I. Davis and A. Burns. *ACM Computing Survey* 43(4), 2011.

[0] *Throughput analysis of synchronous data-flow graphs*. Ghamarian, A.H. et al. Application of Concurrency to System Design. IEEE, 2006

[0] *Latency minimization for synchronous data flow graphs*. A. H. Ghamarian, et al. Conference on Digital System Design Architectures, Methods and Tools. Euromicro, 2007.

[0] *Minimal memory schedules for data-flow networks*. M. Cubric and P. Panangaden. International Conference on Concurrency Theory. Springer, 1993.

[0] *Looped schedules for dataflow descriptions of multirate signal processing algorithms*. S. S. Bhattacharyya and E. A. Lee. Journal of Formal Methods in System Design. Kluwer, 1994.

[0] *Affine data-flow graphs for the synthesis of hard real-time applications*. A. Bouakaz, J.-P. Talpin, and J. Vitek. International Conference on Application of Concurrency to System Design. IEEE Press, 2012.

[0] *Hard-real-time scheduling of data-dependent tasks in embedded streaming applications*. M. Bamakhrama and T. Stefanov. International Conference on Embedded Software. ACM, 2011.

[0] *The fast methodology for high-speed SOC simulation.* D. Chiou, et al. International conference on Computer-aided design. IEEE, 2007.

[0] *Using binary translation in event driven simulation for fast and flexible MPSOC simulation.* M. Gligor, N. Fournel, and F. Pétrot. In CODES+ISSS, IEEE, 2009.

hundreds in a comparison between their cycle accurate simulator and their virtual prototyping framework. A factor of the order of 100 times faster than EDA tools is required for virtual prototyping.

In order to speed-up simulation time, the virtual prototype must trade-off with something. Depending upon the application designers goals, one may be interested in trading some loss of accuracy in exchange for simulation speed, which leads to constructing simulation models that focus on some design aspects and provide abstraction of others. A simulation model can provide an abstraction of the simulated hardware in three directions:

- *Computation abstraction*. A hardware component computes a high level function by carrying out a series of small steps executed by composing logical gates. In a virtual prototyping environment, it is often possible to compute the high level function directly by using the available computing resources on the simulation host machine, thus abstracting the hardware function.
- *Communication abstraction*. Hardware components communicate together using some wiring, and some protocol to transmit the data. Simulation of the communication and the particular protocol may be irrelevant for the purpose of virtual prototyping: communication can be abstracted into higher level data transmission functions.
- *Timing Abstraction*. In a cycle accurate simulator, there are multiple simulation tasks, and each task makes some progress on each clock cycle, but this is slowing down the simulation. In a virtual prototyping experiment, one may not need to so precise timing information: coarser time abstractions can be defined allowing for faster simulation.

The cornerstone of a virtual prototyping platform is the component that simulates the processor(s) of the platform, and its associated peripherals. Such simulation can be *static* or *dynamic*.

3.6. Research Objectives

The challenges addressed by team TEA support the claim that sound Cyber-Physical System design (including embedded, reactive, and concurrent systems altogether) should consider (logical, formal) time modelling as a central aspect.

In this aim, architectural specifications found in software engineering are a natural focal point to start from. Architecture descriptions organise a system model into manageable components, establish clear interfaces between them, and help correct integration of these components during system design.

The definition of a formal design methodology to support the heterogeneous modelling of time in architecture descriptions demands the elaboration of sound mathematical foundations and the development of formal calculi methods to instrument them that constitute the research program of team TEA.

3.6.1. Objective n. 1 – Semantics and specification of time in system design

Time systems. To mitigate and generalise algebraic representations of time, we propose to introduce the paradigm of "time system" (type systems to represent time). Just as a type system abstracts data carried along operations in a program, a time system abstracts the causal interaction of that program module or hardware element with its environment, its pre and post conditions, its assumptions and guarantees, either logical or numerical. Instances of the concept of time system we envision are the clock calculi found in data-flow synchronous languages like Signal, Lustre and its different incarnations. All are bound to a particular model of time.

To gain generality and compositionality, we wish to proceed from recent developments on hybrid types [0] (linked to interface and contract theories), refinement types [0], value-dependant type [0] theories, to formally define a time system.

[0] *Hybrid type checking*. K.W. Knowles and C. Flanagan. ACM Transactions on Programming languages and systems, 32(2). ACM, 2010

[0] *Abstract Refinement Types*. N. Vazou, P. Rondon, and R. Jhala. European Symposium on Programming. Springer, 2013.

[0] *Secure distributed programming with value-dependent types*. N. Swamy, et al. International Conference on Functional Programming. Springer, 2011.

The principle of these type systems is to allow data-types inferred in the program with properties, possibly temporal, pertaining, for instance, to the algebraic domain on their value, or any algebraic property related to its computation: effect, memory usage [0], pre-post condition, value-range, cost, speed, time.

In the quest of an appropriate algebra for time, we are studying both the CCSL and PSL standards and, more generally, Kleene algebras [0] which offer greater expressivity in the prospect of timed specification as well as refinement checking and verification [00].

Being grounded on type and domain theories, a time system can naturally be equipped with program analysis techniques based on type inference (for data-type inference) or abstract interpretation (for program properties inference) [0]. We intend to use and learn from existing open-source implementations in this field of research [0] in order to prototype our solution.

Relating time systems. Just as a time system formally represents the timed behaviour of a given component, timing relations (abstraction and refinement) represent interaction among components. Logically, their specification should be the role of a module system, and verifying their conformance that of a module checking algorithm.

Scalability and compositionality dictate the use of assume-guarantee reasoning, as found in interface automata and contract algebra, in order to facilitate composition by behavioural sub-typing, in the spirit of the (static) contract-based formalism proposed by Passerone et al. [00].

To further elaborate a formal verification approach, we will additionally consider notions of refinement calculi based on temporal logic [0], in order to possibly extend our interface and contract theories with liveness properties. The definition of a module/interface for timed architectures should hence proceed directly from the definition of its time system, using mostly existing theoretical results on the matter of module systems, interface and contract theories.

Conformance of time relations. Verification problems encompassing heterogeneously timed specifications are common and of great variety: checking correctness between abstract and concrete time models relates to desynchronisation (from synchrony to asynchrony) and scheduling analysis (from synchrony to hardware). More generally, they can be perceived from heterogeneous timing viewpoints (e.g. mapping a synchronous-time software on a real-time middleware or hardware).

This perspective demands capabilities not only to inject time models one into the other (by abstract interpretation, using refinement calculi), to compare time abstractions one another (using simulation, refinement, bisimulation, equivalence relations) but also to prove more specific properties (synchronisation, determinism, endochrony).

---

[0] *Region-based memory management*. Tofte, M., Talpin, J.-P. Information and Computation, 132(2). Academic Press, 1997.

[0] *Automated reasoning in Kleene algebra*. P. Höfner and G. Struth. Conference on Automated Reasoning. Springer, 2007.

[0] *Algebraic Verification Method for SEREs Properties via Groebner Bases Approaches*. N. Zhou, J. Wu, X. Gao. Journal of Applied Mathematics. Hindawi, 2013

[0] *From monadic logic to PSL*. M. Y. Vardi. Pillars of Computer Science, 2008.

[0] *Timed polyhedra analysis for synchronous languages*. Besson, F., Jensen, T., Talpin, J.-P. Static Analysis Symposium. Springer, 1999.

[0] The Microsoft F* project, https://research.microsoft.com/en-us/projects/fstar.

[0] *A contract-based formalism for the specification of heterogeneous systems*. L. Benvenistu, A. Ferrari, L. Mangeruca, E. Mazzi, R. Passerone, C. Sofronis. Forum on design languages, 2008

[0] *Moving from Specifications to Contracts in Component-Based Design.* S. Bauer, A. David, R. Hennicker, K. Larsen, A. Legay, U. Nyman, A. Wasowski. Fundamental Aspects in Software Engineering. Springer, 2012

[0] *Refinement Calculus: A Systematic Introduction*. R.J. Back, J. von Wright. Springer, 1998.

In the spirit of our recent work developing an abstract scheduling theory, we want to develop a method of abstract interpretation [0] to reason about the abstraction and refinement of heterogeneous timed specifications in the aim of checking their conformance. A source of inspiration in that prospect is the notion of contract abstraction [0]. To this end, we plan to use SAT-SMT solving techniques to check conformance of abstracted time constraints, in a way which we previously experienced with the automated code generation validation of Polychrony [000].

To check conformance between heterogeneously timed specifications, we will consider variants of the abstract interpretation framework proposed by Bertrane et al. [0] to inject properties from one time domain into another, be it continuous [0] or discrete [0].

This will for instance enable the possibility of verifying cross-domain properties, e.g. cost v.s. power v.s. performance v.s. software mapping. This will allow to formalise intuitions such as that this typical inter-domain constraint: the cost of a system has an impact on the system's controlability; and allow to formally explain why: lower cost means hardware with lower performances, which means longer WCRTs, which means longer end-to-end latency, which may result in a response-time longer than controllability limits. This particular topic (which we could call cross-domain conformance checking) has not been studied in the related literature (on contract-based design, for instance), and could be based on both abstraction techniques, e.g. linear abstractions, or morphisms between domains or even discrete relations, e.g. a simple catalog or "price list" relating price and performance for a data-base of hardware components.

3.6.2. Objective n. 2 – A standard for modelling time in system design

A second objective, to be developed in parallel and synergy to objective n. 1, is the definition of an architecture-specific specification formalism, that would serve as semantic foundation, structure and repository for tooling a component-based design methodology with semantic analysis, to synthesise component interfaces, and formal methods, to verify specified requirements.

In project TEA, it will take form by the definition and tooling of a time annex for the AADL standard, based on the theory developed in objective n. 1. The aim of the AADL time annex is to formalise the logical and physical timing properties of architecture models and represent them as constraints expressed using regular grammars (like in PSL), or using the process calculus of CCSL.

This is an objective reminiscent and in direct application of the principle of time system (objective n.1). We not only want to model time in the heterogeneous logical and physical constituents in an AADL specification, but relate them, and verify the correctness of their composition.

Our aim is to start from the modelling standards AADL and CCSL to define a standard for time in system design. Our contribution will be formalised by a timing annex for the AADL and tools collaboratively developed to support its use. Our first milestone in this prospect is a report [0] of recommendations accepted by the AADL committee. Our next step, the submission of a time annex by team TEA at the SAE consortium, will

[0] *La vérification de programmes par interprétation abstraite*. P. Cousot. Séminaire au Collège de France, 2008.

[0] *Compositional contract abstraction for system design*. A. Benveniste, D. Nickovic, T. Henzinger.

[0] *Efficient deadlock detection for polychronous data-flow specifications*. C. Ngo, J.-P. Talpin, T. Gautier. Electronic System Level Synthesis Conference (ESLSYN'14). IEEE, 2014.

[0] *Formal verification of synchronous data-flow program transformations toward certified compilation*. V.-C. Ngo, J.-P. Talpin, Gautier, P. Le Guernic, L. Besnard. Frontiers of Computer Systems. Springer, 2013.

[0] *Enhancing the Compilation of Synchronous Dataflow Programs with a Combined Numerical-Boolean Abstraction*. P. Feautrier, A. Gamatié and L. Gonnord. Journal of Computing, 1(4). Computer Society of India, 2012.

[0] *Temporal Abstract Domains*. J. Bertrane. International Conference on Engineering of Complex Computer Systems. IEEE, 2011

[0] *Abstract Interpretation of the Physical Inputs of Embedded Programs*. O. Bouissou, M. Martel. Verification, Model Checking, and Abstract Interpretation. LNCS 4905, Springer, 2008

[0] *Proving the Properties of Communicating Imperfectly-Clocked Synchronous Systems*. J. Bertrane. Static Analysis Symposium. Springer, 2006

[0] *"Logically timed specifications in the AADL – Recommendations to the SAE committee on AADL*. L. Besnard, E. Borde, P. Dissaux, T. Gautier, P. Le Guernic, J.-P. Talpin, H. Yu. Inria Technical Report n.446, 2014.

employ the principles exposed in objective n.1 in order to formally define a modular and scalable specification formalism to specify heterogeneous timing constraints in the AADL.

Then, the specification of timing relations between AADL objects will be made explicit by contracts. Together with these contracts, we will then formally define abstraction and refinement relation in order to inject properties assumed by one component into the time model guaranteed by another, and vice versa. Lastly, conformance-checking abstracted contracts will be supported by state-of-the-art verification tools. This all will define a design methodology for time in the AADL, and our very last step will be to tool this methodology and provide a reference implementation.

3.6.3. Objective n. 3 – Applications to real-time scheduling

As a prime application of formal methods for interacting time models, scheduling thousands of program blocks or modules found on modern embedded architecture poses a challenging problem. It simply defies known bounds of complexity theory in the field. It is an issue that requires a particular address, because it would find direct industrial impact in present collaborative projects in which we are involved.

One recent milestone in the prospect of large-scale scheduling is the development of abstract affine scheduling [0]. It consists, first, of approximating threads communication patterns in Safety-Critical Java using cyclo-static data-flow graphs and affine functions. Then, it uses state of the art ILP techniques to find optimal schedules and concretise them as real-time schedules for Safety Critical Java programs [00]

To develop the underlying theory of this promising research topic, we first need to deepen the theoretical foundation to establish links between scheduling analysis and abstraction interpretation [0].

The theory of time system developed in objective n.1 offers the ideal framework to pursue this development. It amounts to representing scheduling constraints, inferred from programs, as types. It allows to formalise the target time model of the scheduler (the architecture, its middle-ware, its real-time system) and defines the basic concepts to verify assumptions made in one with promises offered by the other: contract verification or, in this case, synthesis. Objective n.3 is hence defined as a direct application of objective n.1.

3.6.4. Objective n. 4 – Applications to virtual prototyping

A solution usually adopted to handle time in virtual prototyping is to manage hierarchical time scales, use component abstractions where possible to gain performance, use refinement to gain accuracy where needed. Localised time abstraction may not only yield faster simulation, but facilitate also verification and synthesis (e.g. synchronous abstractions of physically distributed systems). Such an approach requires computations and communications to be harmoniously discretised and abstracted from originally heterogeneous viewpoints onto a structuring, articulating, pivot model, for concerted reasoning about time and scheduling of events in a way that ensures global system specification correctness.

Just as model checking usually employs goal-directed abstraction techniques, in order to approximate parts of the model that are not in the path of the property to check, we plan to equivalently define, possibly semi-automate, abstraction techniques to approximate the time model of system components that do not directly influence timing properties to evaluate.

In the short term these component models could be based on libraries of predefined models of different levels of abstractions. Such abstractions are common in large programming workbench for hardware modelling, such as SystemC, but less so, because of the engineering required, for virtual prototyping platforms. Additionally, the level of abstraction required to simulate components could simply (and best) be specified manually by annotating the architecture specification.

[0] *Buffer minimization in earliest-deadline first scheduling of dataflow graphs*. A. Bouakaz and J.-P. Talpin. Conference on Languages, Compilers and Tools for Embedded Systems. ACM, June 2013.

[0] *Affine data-flow graphs for the synthesis of hard real-time applications*. A. Bouakaz, J.-P. Talpin, and J. Vitek. Application of Concurrency to System Design. IEEE Press, June 2012.

[0] *Design of Safety-Critical Java Level 1 Applications Using Affine Abstract Clocks*. A. Bouakaz and J.-P. Talpin. International Workshop on Software and Compilers for Embedded Systems. ACM, June 2013.

[0] *Abstraction-Refinement for Priority-Driven Scheduling of Static Dataflow Graphs*. Submitted for publication, 2014.

The approach of team TEA provides an additional ingredient in the form of rich component interfaces. It therefore dictates to further investigate the combined use of conventional virtual prototyping libraries, defined as executable abstractions of real hardware, with executable component simulators synthesised from rich interface specifications (using, e.g., conventional compiling techniques used for synchronous programs).

Just as virtual integration consists of synthesising the verification model of an architecture specification, virtual prototyping can be seen as synthesising an executable simulator from a model in, e.g., the spirit of the A-350 DMS case study that was realised by team ESPRESSO in the frame of Artemisia project CESAR [0].

[0] *System-level co-simulation of integrated avionics using polychrony*. Yu, H., Ma, Y., Glouche, Y., Talpin, J.-P., Besnard, L., Gautier, T., Le Guernic, P., Toom, A., and Laurent, O. ACM Symposium on Applied Computing. ACM, 2011.

TEMPO Team

3. Research Program

3.1. Cyber Physical Systems

The development of complex embedded systems platforms requires putting together many hardware components, processor cores, application specific co-processors, bus architectures, peripherals, etc. The hardware platform of a project is seldom entirely new. In fact, in most cases, 80 percent of the hardware components are re-used from previous projects or simply are COTS (Commercial Off-The-Shelf) components. There is no need to simulate in great detail these already proven components, whereas there is a need to run fast simulation of the software using these components.

These requirements call for an integrated, modular simulation environment where already proven components can be simulated quickly, (possibly including real hardware in the loop), new components under design can be tested more thoroughly, and the software can be tested on the complete platform with reasonable speed.

Modularity and fast prototyping also have become important aspects of simulation frameworks, for investigating alternative designs with easier re-use and integration of third party components. The project aims at developing such a rapid prototyping, modular simulation platform, combining new hardware components modeling, verification techniques, fast software simulation for proven components, capable of running the real embedded software application without any change.

To fully simulate a complete hardware platform, one must simulate the processors and co-processors, together with the peripherals such as network controllers, graphics controllers, USB controllers, etc. A commonly used solution is the combination of some ISS (Instruction Set Simulator) connected to a Hardware Description Language (HDL) simulator, in a co-simulation environment such as [12], [13]. Some communication and synchronization must be designed and maintained between the two using some inter-process communication (IPC), which slows down the process.

The idea we pursue is to combine hardware modeling and fast simulation into a fully integrated, software based simulation environment, which uses a single simulation loop thanks to Transaction Level Modeling (TLM) [3] combined with a new ISS technology designed specifically to fit within the TLM environment.

The most challenging way to enhance simulation speed is to simulate the processors. Processor simulation is achieved with Instruction Set Simulation (ISS). There are several alternatives to achieve such simulation. In *interpretive simulation*, each instruction of the target program is fetched from memory, decoded, and executed. This method is flexible and easy to implement, but the simulation speed is slow as it wastes a lot of time in decoding. Interpretive simulation is used in Simplescalar [2]. Another technique to implement a fast ISS is *dynamic translation* [8], [4] which has been favored by many implementors [18], [19], [20], [14] in the past decade.

There are many ways of translating binary code into cached data, which each come at a price, with different trade-offs between the translation time and the obtained speed up on cache execution. Also, simulation speed-ups usually don't come for free: most of time there is a trade-off between accuracy and speed. There are two well known variants of the dynamic translation technology: the target code is translated either directly into machine code for the simulation host, or into an intermediate representation, independent from the host machine, that makes it possible to execute the code with faster speed. A challenge in the development of high performance simulators is to maintain simultaneously fast speed and simulation accuracy. In the TEMPO project, we expect to develop a dynamic translation technology satisfying the following additional objectives:

- provide different levels of translation with different degrees of accuracy so that users can choose between accurate and slow (for debugging) or less accurate but fast simulation.
- to take advantage of multi-processor simulation hosts to parallelize the simulation;
- to define intermediate representations of programs that optimize the simulation speed and possibly provide a more convenient format for studying properties of the simulated programs.

Another objective of the TEMPO simulation is to extract information from the simulated applications in order to prove system properties. One can use model based tools to generate tests that can be run on the simulator to check whether the test fails or not on the real application. The project is considering an approach as illustrated in Figure 1

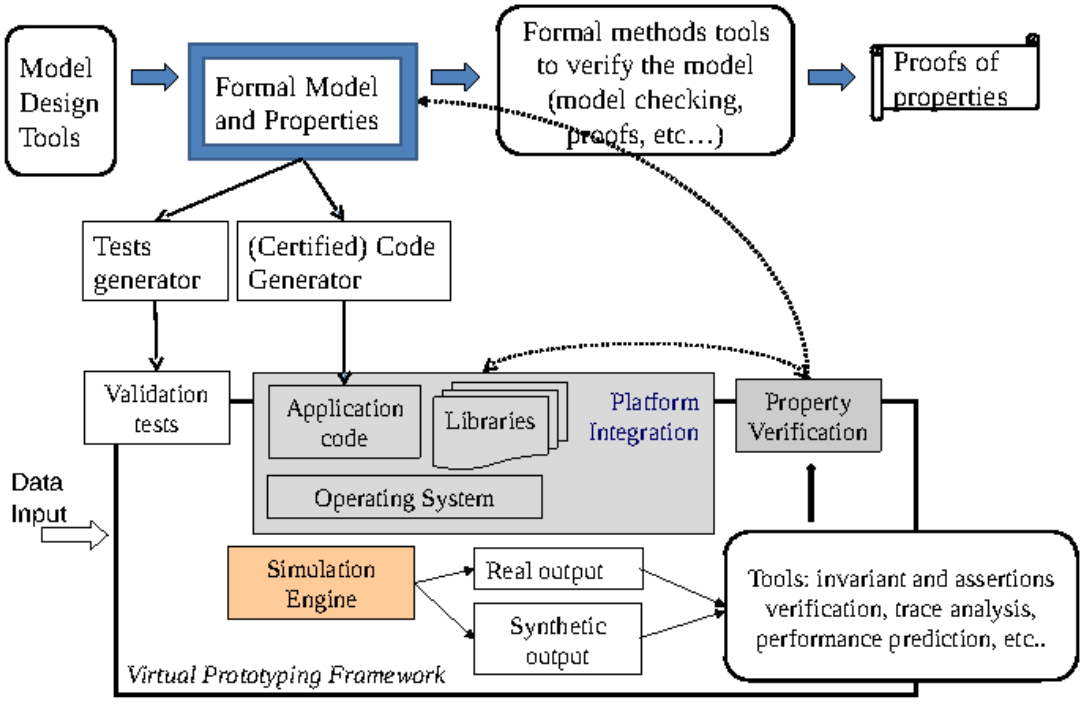


Figure 1. Proposed development methodology

Thus, it is also a goal of TEMPO activities to use such formal methods tools to detect failures, either by generating tests, or by using formal methods tools to analyze the results of simulation sessions.

3.2. Verification of Embedded Systems Properties

Since last decade, we have witnessed rapid development in embedded system domain. More and more state-of-the-art embedded systems adopt the heterogeneous multi-processor platform rather than the platform with single core. To achieve better quality and performance, the design paradigm shift from simple control system to complex heterogeneous Cyber-Physical Systems (CPS) is gaining more interests. Increasing complexity coupled with time-to-market pressure create a critical need to validate heterogeneous embedded system designs. The functional validation is thus widely acknowledged as a major bottleneck in embedded system design. To guarantee the reliability of heterogeneous embedded systems, up to 70% of the overall design time and resources are spent on functional validation.

From the verification point of view, the major objective of this project is to reduce the overall validation efforts in the top-down design flow of embedded system design using the high-level specifications. In this project, we plan to address the following three major problems:

- **Formal modeling of high-level specifications**. We want to investigate how to model heterogeneous systems with multiple models of computation (MoC) and how to extract the formal models from system-level specifications to enable automated analysis.
- **Efficient validation of system-level specifications with minimum effort**. The idea here is to investigate the automated directed test generation from high-level specification validation and explore various approaches and techniques to further reduce the directed test generation time (eliminate redudant tests).
- **Consistency checking between different abstraction layers**. We also want to explore the possibility of reusing high-level validation efforts for low-level implementation validation as well as to check the consistency between different abstraction layers.

In conclusion, this project targets to improve the effectiveness and efficiency of functional validation of heterogeneous embedded systems. We believe that our approaches can not only enhance the reliability of heterogeneous embedded systems, but also reduce the time-to-market.

TOCCATA Project-Team

3. Research Program

3.1. Introduction

In the former *ProVal* project, we have been working on the design of methods and tools for deductive verification of programs. One of our originalities is our ability to conduct proofs by using automatic provers and proof assistants at the same time, depending on the difficulty of the program, and specifically the difficulty of each particular verification condition. We thus believe that we are in a good position to propose a bridge between the two families of approaches of deductive verification presented above. This is a new goal of the team: we want to provide methods and tools for deductive program verification that can offer both a high amount of proof automation and a high guarantee of validity. Toward this objective, we propose a new axis of research: to develop certified tools, i.e. analysis tools that are themselves formally proved correct.

As mentioned above, some of the members of the team have an internationally-recognized expertise on deductive program verification involving floating-point computation [6], including both interactive proving and automated solving [10]. Indeed we noticed that the verification of numerical programs is a representative case that can benefit a lot from combining automatic and interactive theorem proving [64], [5]. This motivated our research on the formal verification of numerical programs.

Moreover, we continue the fundamental studies we conducted in the past concerning deductive program verification in general. This is why our detailed scientific programme is structured into three themes:

1. Formally Verified Programs,
2. Certified Tools,
3. Numerical Programs.

3.2. Formally Verified Programs

Formal program verification is a research theme that builds upon our expertise in the development of methods and tools for proving programs, from source codes annotated with specifications to proofs. In the past, we tackled programs written in mainstream programming languages, with the system *Why3* and the front-ends *Krakatoa* for Java source code, and *Frama-C/Jessie* for C code. However, Java and C programming languages were designed a long time ago, and certainly not with the objective of formal verification in mind. This raises a lot of difficulties when designing specification languages on top of them, and verification condition generators to analyze them. On the other hand, we designed and/or used the *Coq* and *Why3* languages and tools for performing deductive verification, but those were not designed as programming languages that can be compiled into executable programs.

Thus, a new axis of research we propose is the design of an environment that is aimed to both programming and proving, hence that will allow to develop correct-by-construction programs. To achieve this goal, there are two major axes of theoretical research that needs to be conducted, concerning, on the one hand, methods required to support genericity and reusability of verified components, and, on the other hand, the automation of the proof of the verification conditions that will be generated.

3.2.1. Genericity and Reusability of Verified Components

A central ingredient for the success of deductive approaches in program verification is the ability to reuse components that are already proved. This is the only way to scale the deductive approach up to programs of larger size. As for programming languages, a key aspect that allow reusability is *genericity*. In programming languages, genericity typically means parametricity with respect to data types, e.g. *polymorphic types* in functional languages like ML, or *generic classes* in object-oriented languages. Such genericity features are essential for the design of standard libraries of data structures such as search trees, hash tables, etc. or libraries of standard algorithms such as for searching, sorting.

In the context of deductive program verification, designing reusable libraries also requires designing of *generic specifications* which typically involve parametricity not only with respect to data types but also with respect to other program components. For example, a generic component for sorting an array needs to be parametrized by the type of data in the array but also by the comparison function that will be used. This comparison function is thus another program component that is a parameter of the sorting component. For this parametric component, one needs to specify some requirements, at the logical level (such as being a total ordering relation), but also at the program execution level (like being *side-effect free*, i.e. comparing of data should not modify the data). Typically such a specification may require *higher-order* logic.

Another central feature that is needed to design libraries of data structures is the notion of data invariants. For example, for a component providing generic search trees of reasonable efficiency, one would require the trees to remain well-balanced, over all the life time of a program.

This is why the design of reusable verified components requires advanced features, such as *higher-order specifications and programs*, *effect polymorphism* and *specification of data invariants*. Combining such features is considered as an important challenge in the current state of the art (see e.g. [98]). The well-known proposals for solving it include *Separation logic* [121], *implicit dynamic frames* [118], and *considerate reasoning* [120]. Part of our recent research activities were aimed at solving this challenge: first at the level of specifications, e.g. we proposed generic specification constructs upon Java [122] or a system of theory cloning in our system *Why3* [2]; second at the level of programs, which mainly aims at controlling side-effects to avoid unexpected breaking of data invariants, thanks to advanced type checking: approaches based on *memory regions*, *linearity* and *capability-based* type systems [72], [96], [51].

A concrete challenge that should be solved in the future is: what additional constructions should we provide in a specification language like ACSL for C, in order to support modular development of reusable software components? In particular, what would be an adequate notion of module, that would provide a good notion of abstraction, both at the level of program components and at the level of specification components?

3.2.2. Automated Deduction for Program Verification

Verifying that a program meets formal specifications typically amounts to generating *verification conditions* e.g. using a weakest precondition calculus. These verification conditions are purely logical formulas—typically in first-order logic and involving arithmetic in integers or real numbers—that should be checked to be true. This can be done using either automatic provers or interactive proof assistants. Automatic provers do not need user interaction, but may run forever or give no conclusive answer.

There are several important issues to tackle. Of course, the main general objective is to improve automation as much as possible. We continue our efforts around our own automatic prover *Alt-Ergo* towards more expressivity, efficiency, and usability, in the context of program verification. More expressivity means that the prover should better support the various theories that we use for modeling. Toward this direction, we aim at designing specialized proof search strategies in *Alt-Ergo*, directed by rewriting rules, in the spirit of what we did for the theory of associativity and commutativity [7].

A key challenge also lies in the handling of quantifiers. SMT solvers, including *Alt-Ergo*, deal with quantifiers with a somewhat ad-hoc mechanism of heuristic instantiation of quantified hypotheses using the so-called *triggers* that can be given by hand [83], [84]. This is completely different from resolution-based provers of the TPTP category (E-prover, Vampire, etc.) which use unification to apply quantified premises. A challenge is thus to find the best way to combine these two different approaches of quantifiers. Another challenge is to add some support for higher-order functions and predicates in this SMT context, since as said above, reusable verified components will require higher-order specifications. A few solutions have been proposed, essentially based on encoding of higher-order goals into first-order goals [96].

Generally speaking, there are several theories, interesting for program verification, that we would like to add as built-in decision procedures in an SMT context. First, although there already exist decision procedures for variants of bit-vectors, they are not complete enough to support what is needed to reason on programs that manipulate data at the bit-level, in particular if conversions from bit-vectors to integers or floating-point

numbers are involved [114]. Regarding floating-point numbers, an important challenge is to integrate in an SMT context a decision procedure like the one implemented in our tool *Gappa*.

Another goal is to improve the feedback given by automatic provers: failed proof attempts should be turned into potential counterexamples, so as to help debugging programs or specifications. A pragmatic goal would be to allow cooperation with other verification techniques. For instance, testing could be performed on unproved goals. Regarding this cooperation objective, an important goal is a deeper integration of automated procedures in interactive proofs, like it already exists in Isabelle [70]. We now have a *Why3* tactic in *Coq* that we plan to improve.

3.2.3. An Environment for Both Programming and Proving

As said before, a new axis of research we follow is the design of a language and an environment for both programming and proving. We believe that this will be a fruitful approach for designing highly trustable software. This is a similar goal as projects Plaid, Trellys, ATS, or Guru, mentioned above.

The basis of this research direction is the *Why3* system, which is in fact a reimplementation from scratch of the former *Why* tool, that we started in January 2011. This new system supports our research at various levels. It is already used as an intermediate language for deductive verification.

The next step for us is to develop its use as a true programming language. Our objective is to propose a language where programs could be both executed (e.g. thanks to a compiler to, say, *OCaml*) and proved correct. The language would basically be purely applicative (i.e. without side-effects, e.g. close to ML) but incorporating specifications in its core. There are, however, some programs (e.g. some clever algorithms) where a bit of imperative programming is desirable. Thus, we want to allow some form of imperative features, but in a very controlled way: it should provide a strict form of imperative programming that is clearly more amenable to proof, in particular dealing with data invariants on complex data structures.

As already said before, reusability is a key issue. Our language should propose some form of modules with interfaces abstracting away implementation details. Our plan is to reuse the known ideas of *data refinement* [110] that was the foundation of the success of the B method. But our language will be less constrained than what is usually the case in such a context, in particular regarding the possibility of sharing data, and the constraints on composition of modules, there will be a need for advanced type systems like those based on regions and permissions.

The development of such a language will be the basis of the new theme regarding the development of certified tools, that is detailed in Section 3.3 below.

3.2.4. Extra Exploratory Axes of Research

Concerning formal verification of programs, there are a few extra exploratory topics that we plan to explore.

Concurrent Programming So far, we only investigated the verification of sequential programs. However, given the spreading of multi-core architectures nowadays, it becomes important to be able to verify concurrent programs. This is known to be a major challenge. We plan to investigate this direction, but in a very careful way. We believe that the verification of concurrent programs should be done only under restrictive conditions on the possible interleaving of processes. In particular, the access and modification of shared data should be constrained by the programming paradigm, to allow reasonable formal specifications. In this matter, the issues are close to the ones about sharing data between components in sequential programs, and there are already some successful approaches like separation logic, dynamic frames, regions, and permissions.

Resource Analysis The deductive verification approaches are not necessarily limited to functional behavior of programs. For example, a formal termination proof typically provides a bound on the time complexity of the execution. Thus, it is potentially possible to verify resources consumption in this way, e.g. we could prove WCET (Worst Case Execution Times) of programs. Nowadays, WCET analysis is typically performed by abstract interpretation, and is applied on programs with particular shape (e.g. no unbounded iteration, no recursion). Applying deductive verification techniques in this context could allow to establish good bounds on WCET for more general cases of programs.

Other Programming Paradigms We are interested in the application of deductive methods in other cases than imperative programming à la C, Java or Ada. Indeed, in the recent years, we applied proof techniques to randomized programs [1], to cryptographic programs [50]. We plan to use proof techniques on applications related to databases. We also have plans to support low-level programs such as assembly code [86], [113] and other unstructured programming paradigm. We are also investigating more and more applications of SMT solving, e.g. in model-checking approach (for example in Cubicle [0] [76]) or abstract interpretation techniques (project Cafein, started in 2013) and also for discharging proof obligations coming from other systems like *Atelier B* [109] (project BWare).

3.3. Certified Tools

One of our goals is to guarantee the soundness of the tools we develop. In fact, it goes beyond that; our goal is to promote our future *Why3* environment so that *others* could develop certified tools. Tools like automated provers or program analyzers are good candidate case studies because they are mainly performing symbolic computations, and as such they are usually programmed in a mostly purely functional style.

We conducted several experiments of development of certified software in the past. First, we have a strong expertise in the development of *libraries* in *Coq*: the Coccinelle library [78] formalizing term rewriting systems, the Alea library [1] for the formalization of randomized algorithms, several libraries formalizing floating-point numbers (Floats [60], Gappalib [107], and now Flocq [6] which unifies the formers). Second we conducted the development of a certified decision procedure [103] that corresponds to a core part of *Alt-Ergo*. Third we developed, still in *Coq*, certified verification condition generators, in a first step [94] for a language similar to *Why*, and in a second step [93] for C annotated in ACSL [56], based on the operational semantics formalized in the CompCert certified compiler project [102].

To go further, we have several directions of research in mind.

3.3.1. Formalization of Binders

Using the *Why3* programming language instead of *Coq* allows for more freedom. For example, it should allow one to use a bit of side-effects when the underlying algorithm justifies it (e.g. hash-consing, destructive unification). On the other hand, we will lose some *Coq* features like dependent types that are usually useful when formalizing languages. Among the issues that should be studied, we believe that the question of the formalization of binders is both central and challenging (as exemplified by the POPLmark international challenge [47]).

The support of binders in *Why3* should not be built-in, but should be under the form of a reusable *Why3* library, that should already contain a lot of proved lemmas regarding substitution, alpha-equivalence and such. Of course we plan to build upon the former experiments done for the POPLmark challenge. Although, it is not clear yet that the support of binders only via a library will be satisfactory. We may consider addition of built-in constructs if this shows useful. This could be a form of (restricted) dependent types as in *Coq*, or subset types as in PVS.

3.3.2. Theory Realizations, Certification of Transformations

As an environment for both programming and proving, *Why3* should come with a standard library that includes both verified libraries of programs, but also libraries of specifications (e.g. theories of sets, maps, etc.).

The certification of those *Why3* libraries of specifications should be addressed too. *Why3* libraries for specifying models of programs are commonly expressed using first-order axiomatizations, which have the advantage of being understood by many different provers. However, such style of formalization does not offer strong guarantees of consistency. More generally, the fact that we are calling different kind of provers to discharge our verification conditions raises several challenges for certification: we typically apply various transformations to go from the *Why3* language to those of the provers, and these transformations should be certified too.

[0] http://cubicle.lri.fr/

A first attempt in considering such an issue was done in earlier work [109]. It was proposed to certify the consistency of a library of specification using a so-called *realization*, which amounts to "implementing" the library in a proof assistant like *Coq*. This is an important topic of the ANR project BWare.

3.3.3. Certified Theorem Proving

The goal is to develop *certified* provers, in the sense that they are proved to give a correct answer. This is an important challenge since there have been a significant amount of soundness bugs discovered in the past, in many tools of this kind.

The former work on the certified core of *Alt-Ergo* [103] should be continued to support more features: more theories (full integer arithmetic, real arithmetic, arrays, etc.), quantifiers. Development of a certified prover that supports quantifiers should build upon the previous topic about binders.

In a similar way, the *Gappa* prover, which is specialized to solving constraints on real numbers and floating-point numbers, should be certified too. However, for very complex decision procedures, developing a certified proof search might be too ambitious. Instead, the idea is to ask *Gappa* to produce Coq proofs on a per-goal basis, so as to check *a posteriori* the soundness of its result on the given instance. More generally, we can have *Gappa* produce traces of its execution that can later be processed by a certified trace checker. This approach was used in the past for certified proofs of termination of rewriting systems [79], and it was also used internally in CompCert for several passes of compilation [102].

3.3.4. Certified VC Generation

The other kind of tools that we would like to certify are the VC generators. This is a continuation of the work on developing in *Coq* a certified VC generator for C code annotated in ACSL [93]. We develop such a generator in *Why3* instead of *Coq* [105]. As before, this builds upon a formalization of binders. There are various kinds of VC generators that are interesting. A generator for a simple language in the style of those of *Why3* is a first step. Other interesting cases are: a generator implementing the so-called *fast weakest preconditions* [99], and a generator for unstructured programs like assembly, that would operate on an arbitrary control-flow graph.

On a longer term, we wish to be able to certify advanced verification methods like those involving refinement, alias control, regions, permissions, etc.

An interesting question is how one could certify a VC generator that involves a highly expressive logic, like higher-order logic, as it is the case of the *CFML* method [73] which allows one to use the whole *Coq* language to specify the expected behavior. One challenging aspect of such a certification is that a tool that produces *Coq* definitions, including inductive definitions and module definitions, cannot be directly proved correct in *Coq*, because inductive definitions and module definitions cannot be generated through the evaluation of *Coq* definitions. Therefore, it seems necessary to involve, in a way or another, a "deep embedding", that is, a formalization of *Coq* in *Coq*, possibly by reusing the deep embedding developed by B. Barras [53].

3.4. Numerical Programs

In recent years, we demonstrated our capability towards specifying and proving properties of floating-point programs, properties which are both complex and precise about the behavior of those programs: see the publications [67], [123], [62], [117], [66], [61], [108], [106] as well as the web galleries of certified programs at our Web page [0], the Hisseo project [0], S. Boldo's page [0], and industrial case studies in the U3CAT ANR project. The ability to express such complex properties comes from models developed in *Coq* [6]. The ability to combine proof by reasoning and proof by computation is a key aspect when dealing with floating-point programs. Such a modeling provides a safe basis when dealing with C source code [5]. However, the proofs can get difficult even on short programs. To build these proofs, some automation is needed. It can be obtained by combining SMT solvers and *Gappa* [64], [82], [46], [10]. Finally, the precision of the verification is obtained thanks to precise models of floating-point computations, taking into account the peculiarities of the architecture (e.g., x87 80-bit floating-point unit) and also the compiler optimizations [68], [113].

[0] http://toccata.lri.fr/gallery/index.en.html
[0] http://hisseo.saclay.inria.fr/
[0] http://www.lri.fr/~sboldo/research.html

The directions of research concerning floating-point programs that we pursue are the following.

3.4.1. Making Formal Verification of Floating-point Programs Easier

A first goal is to ease the formal verification of floating-point programs: the primary objective is still to improve the scope and efficiency of our methods, so as to ease further the verification of numerical programs. The ongoing development of the Flocq library continues towards the formalization of bit-level manipulations and also of exceptional values (e.g. infinities). We believe that good candidates for applications of our techniques are advanced algorithms to compute efficiently with floats, which operate at the bit-level. The formalization of real numbers is being revamped too: higher-level numerical algorithms are usually built on some mathematical properties (e.g. computable approximations of ideal approximations), which then have to be proved during the formal verification of these algorithms.

Easing the verification of numerical programs also implies more automation. SMT solvers are generic provers well-suited for automatically discharging verification conditions, but they appear to loose their effectiveness when floating-point arithmetic is involved [77]. Our goal is to improve the arithmetic theories of *Alt-Ergo*, so that they support floating-point arithmetic along their other theories, if possible by reusing the heuristics developed for *Gappa*.

3.4.2. Continuous Quantities, Numerical Analysis

Our goal is to handle floating-point programs that are related to continuous quantities. This includes numerical analysis programs we have already worked on [63], [62], [4]. But our work is only a beginning: we were able to solve the difficulties to prove one particular scheme for one particular partial differential equation. We need to be able to easily prove other programs of this kind. This requires new results that handle generic schemes and many partial differential equations. The idea is to design a toolbox to prove these programs with as much automation as possible. We wish this could be used by numerical analysts that are not or hardly familiar with formal methods, but are nevertheless interested in the formal correctness of their schemes and their programs.

Another very interesting kind of programs (especially for industrial developers) are those based on *hybrid* systems, that is where both discrete and continuous quantities are involved. This is a longer-term goal, but we may try to go towards this direction. A first problem is to be able to specify hybrid systems: what are they exactly expected to do? Correctness usually means not going into a forbidden state but we may want additional behavioral properties. A second problem is the interface with continuous systems, such as sensors. How can we describe their behavior? Can we be sure that the formal specification fits? We may think about Ariane V where one piece of code was shamelessly reused from Ariane IV. Ensuring that such a reuse is allowed requires to correctly specify the input ranges and bandwidths of physical sensors.

Studying hybrid systems is among the goals of the new ANR project Cafein.

3.4.3. Certification of Floating-point Analyses

In coordination with our second theme, another objective is to port the kernel of *Gappa* into either *Coq* or *Why3*, and then extract a certified executable. Rather than verifying the results of the tool *a posteriori* with a proof checker, they would then be certified *a priori*. This would simplify the inner workings of *Gappa*, help to support new features (e.g. linear arithmetic, elementary functions), and make it scale better to larger formulas, since the tool would no longer need to carry certificates along its computations. Overall the tool would then be able to tackle a wider range of verification conditions.

An ultimate goal would be to develop the decision procedure for floating-point computations, for SMT context, that is mentioned in Section 3.2.2 , directly as a certified program in *Coq* or *Why3*.

VEGAS Project-Team (section vide)

VERIDIS Project-Team

3. Research Program

3.1. Automated and Interactive Theorem Proving

The VeriDis team unites experts in techniques and tools for interactive and automated verification, and specialists in methods and formalisms designed for developing concurrent and distributed systems and algorithms that are firmly grounded on precise mathematical and semantical abstractions. Our common objective is to advance the state of the art in interactive and automatic deduction techniques, and their combinations, resulting in powerful tools for the computer-aided verification of distributed systems and protocols. Our techniques and tools support sound methods for the development of trustworthy distributed systems that scale to algorithms relevant for practical applications.

VeriDis members from Saarbrücken are developing SPASS [10], one of the leading automated theorem provers for first-order logic based on the superposition calculus [46]. Recent extensions to the system include the integration of dedicated reasoning procedures for specific theories, such as linear arithmetic [56], [45], that are ubiquitous in the verification of systems and algorithms. The group also studies general frameworks for the combination of theories such as the locality principle [57] and automated reasoning mechanisms these induce. Finally, members of the group design effective quantifier elimination methods and decision procedures for algebraic theories, supported by their efficient implementation in the Redlog system [4].

In a complementary approach to automated deduction, VeriDis members from Nancy develop veriT [1], an SMT (Satisfiability Modulo Theories [48]) solver that combines decision procedures for different fragments of first-order logic and that integrates an automatic theorem prover for full first-order logic. The veriT solver is designed to produce detailed proofs; this makes it particularly suitable as a component of a robust cooperation of deduction tools.

We rely on interactive theorem provers for reasoning about specifications at a high level of abstraction. Members of VeriDis have ample experience in the specification and subsequent machine-assisted, interactive verification of algorithms. In particular, we participate in a project at the joint MSR-Inria Centre in Saclay on the development of methods and tools for the formal proof of TLA$^+$ [52] specifications. Our prover relies on a declarative proof language, and we contribute several automatic backends [3].

3.2. Formal Methods for Developing Algorithms and Systems

Powerful theorem provers are not a panacea for system verification: they support sound methodologies for modeling and verifying systems. In this respect, members of VeriDis have gained expertise and recognition in making contributions to formal methods for concurrent and distributed algorithms and systems [2], [9], and in applying them to concrete use cases. In particular, the concept of *refinement* [44], [47], [54] in state-based modeling formalisms is central to our approach. Its basic idea is to present an algorithm or implementation through a series of models, starting from a high-level description that precisely states the problem, and gradually adding details in intermediate models. An important goal in designing such methods is to establish precise proof obligations that can be discharged to a high degree by automatic tools. This requires taking into account specific characteristics of certain classes of systems and tailoring the model to concrete computational models. Our research in this area is supported by carrying out case studies for academic and industrial developments. This activity benefits from and influences the development of our proof tools.

Our vision for the integration of our expertise can be resumed as follows. Based on our experience and related work on specification languages, logical frameworks, and automatic theorem proving tools, we develop an approach that is suited for specification, interactive theorem proving, and for eventual automated analysis and verification, possibly through appropriate translation methods. While specifications are developed by users inside our framework, they are analyzed for errors by our SMT based verification tools. Eventually, properties are proved by a combination of interactive and automatic theorem proving tools.

Today, the formal verification of a new algorithm is typically the subject of a PhD thesis, if it is addressed at all. This situation is not sustainable given the move towards more and more parallelism in mainstream systems: algorithm developers and system designers must be able to productively use verification tools for validating their algorithms and implementations. On a high level, the goal of VeriDis is to make formal verification standard practice for the development of distributed algorithms and systems, just as symbolic model checking has become commonplace in the development of embedded systems and as security analysis for cryptographic protocols is becoming standard practice today. Although the fundamental problems in distributed programming, such as mutual exclusion, leader election, group membership or consensus, are well-known, they pose new challenges in the context of modern system paradigms, including ad-hoc and overlay networks or peer-to-peer systems, and they must be integrated for concrete applications.

APICS Project-Team

3. Research Program

3.1. Introduction

Within the extensive field of inverse problems, much of the research by Apics deals with reconstructing solutions of classical elliptic PDEs from their boundary behavior. Perhaps the simplest example lies with harmonic identification of a stable linear dynamical system: the transfer-function f can be evaluated at a point $i\omega$ of the imaginary axis from the response to a periodic input at frequency ω. Since f is holomorphic in the right half-plane, it satisfies there the Cauchy-Riemann equation $\overline{\partial} f = 0$, and recovering f amounts to solve a Dirichlet problem which can be done in principle using, *e.g.* the Cauchy formula.

Practice is not nearly as simple, for f is only measured pointwise in the pass-band of the system which makes the problem ill-posed [72]. Moreover, the transfer function is usually sought in specific form, displaying the necessary physical parameters for control and design. For instance if f is rational of degree n, then $\overline{\partial} f = \sum_1^n a_j \delta_{z_j}$ where the z_j are its poles and δ_{z_j} is a Dirac unit mass at z_j. Thus, to find the domain of holomorphy (*i.e.* to locate the z_j) amounts to solve a (degenerate) free-boundary inverse problem, this time on the left half-plane. To address such questions, the team has developed a two-step approach as follows.

> **Step 1:** To determine a complete model, that is, one which is defined at every frequency, in a sufficiently versatile function class (*e.g.* Hardy spaces). This ill-posed issue requires regularization, for instance constraints on the behavior at non-measured frequencies.
>
> **Step 2:** To compute a reduced order model. This typically consists of rational approximation of the complete model obtained in step 1, or phase-shift thereof to account for delays. We emphasize that deriving a complete model in step 1 is crucial to achieve stability of the reduced model in step 2.

Step 1 relates to extremal problems and analytic operator theory, see Section 3.3.1 . Step 2 involves optimization, and some Schur analysis to parametrize transfer matrices of given Mc-Millan degree when dealing with systems having several inputs and outputs, see Section 3.3.2.2 . It also makes contact with the topology of rational functions, in particular to count critical points and to derive bounds, see Section 3.3.2 . Step 2 raises further issues in approximation theory regarding the rate of convergence and the extent to which singularities of the approximant (*i.e.* its poles) tend to singularities of the approximated function; this is where logarithmic potential theory becomes instrumental, see Section 3.3.3 .

Applying a realization procedure to the result of step 2 yields an identification procedure from incomplete frequency data which was first demonstrated in [78] to tune resonant microwave filters. Harmonic identification of nonlinear systems around a stable equilibrium can also be envisaged by combining the previous steps with exact linearization techniques from [36].

A similar path can be taken to approach design problems in the frequency domain, replacing the measured behavior by some desired behavior. However, describing achievable responses in terms of the design parameters is often cumbersome, and most constructive techniques rely on specific criteria adapted to the physics of the problem. This is especially true of filters, the design of which traditionally appeals to polynomial extremal problems [74], [59]. Apics contributed to this area the use of Zolotarev-like problems for multi-band synthesis, although we presently favor interpolation techniques in which parameters arise in a more transparent manner, see Section 3.2.2 .

The previous example of harmonic identification quickly suggests a generalization of itself. Indeed, on identifying $\mathbb{C}$ with $\mathbb{R}^2$, holomorphic functions become conjugate-gradients of harmonic functions, so that harmonic identification is, after all, a special case of a classical issue: to recover a harmonic function on a domain from partial knowledge of the Dirichlet-Neumann data; when the portion of boundary where data are not available is itself unknown, we meet a free boundary problem. This framework for 2-D non-destructive control was first advocated in [64] and subsequently received considerable attention. It makes clear how to

state similar problems in higher dimensions and for more general operators than the Laplacian, provided solutions are essentially determined by the trace of their gradient on part of the boundary which is the case for elliptic equations [0] [25], [83]. Such questions are particular instances of the so-called inverse potential problem, where a measure μ has to be recovered from the knowledge of the gradient of its potential (*i.e.*, the field) on part of a hypersurface (a curve in 2-D) encompassing the support of μ. For Laplace's operator, potentials are logarithmic in 2-D and Newtonian in higher dimensions. For elliptic operators with non constant coefficients, the potential depends on the form of fundamental solutions and is less manageable because it is no longer of convolution type. Nevertheless it is a useful concept bringing perspective on how problems could be raised and solved, using tools from harmonic analysis.

Inverse potential problems are severely indeterminate because infinitely many measures within an open set produce the same field outside this set; this phenomenon is called *balayage* [71]. In the two steps approach previously described , we implicitly removed this indeterminacy by requiring in step 1 that the measure be supported on the boundary (because we seek a function holomorphic throughout the right half space), and by requiring in step 2 that the measure be discrete in the left half-plane. The discreteness assumption also prevails in 3-D inverse source problems, see Section 4.2 . Conditions that ensure uniqueness of the solution to the inverse potential problem are part of the so-called regularizing assumptions which are needed in each case to derive efficient algorithms.

To recap, the gist of our approach is to approximate boundary data by (boundary traces of) fields arising from potentials of measures with specific support. Note that it is different from standard approaches to inverse problems, where descent algorithms are applied to integration schemes of the direct problem; in such methods, it is the equation which gets approximated (in fact: discretized).

Along these lines, Apics advocates the use of steps 1 and 2 above, along with some singularity analysis, to approach issues of nondestructive control in 2-D and 3-D [43] [5], [2]. The team is currently engaged in two kinds of generalizations, to be described further in Section 3.2.1 . The first deals with non-constant conductivities in 2-D, where Cauchy-Riemann equations characterizing holomorphic functions are replaced by conjugate Beltrami equations characterizing pseudo-holomorphic functions; next in line are 3-D situations that we begin to consider, see Sections 6.2 and 4.4 . There, we seek applications to inverse free boundary problems such as plasma confinement in the vessel of a tokamak, or inverse conductivity problems like those arising in impedance tomography. The second generalization lies with inverse source problems for the Laplace equation in 3-D, where holomorphic functions are replaced by harmonic gradients; applications are to EEG/MEG and inverse magnetization problems in paleomagnetism, see Section 4.2 .

The approximation-theoretic tools developed by Apics to handle issues mentioned so far are outlined in Section 3.3 . In Section 3.2 to come, we describe in more detail which problems are considered and which applications are targeted.

3.2. Range of inverse problems

3.2.1. Elliptic partial differential equations (PDE)

Participants: Laurent Baratchart, Sylvain Chevillard, Juliette Leblond, Christos Papageorgakis, Dmitry Ponomarev.

By standard properties of conjugate differentials, reconstructing Dirichlet-Neumann boundary conditions for a function harmonic in a plane domain, when these boundary conditions are known already on a subset E of the boundary, is equivalent to recover a holomorphic function in the domain from its boundary values on E. This is the problem raised on the half-plane in step 1 of Section 3.1 . It makes good sense in holomorphic Hardy spaces where functions are entirely determined by their values on boundary subsets of positive linear

[0] There is a subtle difference here between dimension 2 and higher. Indeed, a function holomorphic on a plane domain is defined by its non-tangential limit on a boundary subset of positive linear measure, but there are non-constant harmonic functions in the 3-D ball, C^1 up to the boundary sphere, yet having vanishing gradient on a subset of positive measure of the sphere. Such a "bad" subset, however, cannot have interior points on the sphere.

measure, which is the framework for Problem (P) that we set up in Section 3.3.1 . Such issues naturally arise in nondestructive testing of 2-D (or 3-D cylindrical) materials from partial electrical measurements on the boundary. For instance, the ratio between the tangential and the normal currents (the so-called Robin coefficient) tells one about corrosion of the material. Thus, solving Problem (P) where ψ is chosen to be the response of some uncorroded piece with identical shape yields non destructive testing of a potentially corroded piece of material, part of which is inaccessible to measurements. This was an initial application of holomorphic extremal problems to non-destructive control [56], [60].

Another application by the team deals with non-constant conductivity over a doubly connected domain, the set E being now the outer boundary. Measuring Dirichlet-Neumann data on E, one wants to recover level lines of the solution to a conductivity equation, which is a so-called free boundary inverse problem. For this, given a closed curve inside the domain, we first quantify how constant the solution on this curve. To this effect, we state and solve an analog of Problem (P), where the constraint bears on the real part of the function on the curve (it should be close to a constant there), in a Hardy space of a conjugate Beltrami equation, of which the considered conductivity equation is the compatibility condition (just like the Laplace equation is the compatibility condition of the Cauchy-Riemann system). Subsequently, a descent algorithm on the curve leads one to improve the initial guess. For example, when the domain is regarded as separating the edge of a tokamak's vessel from the plasma (rotational symmetry makes this a 2-D situation), this method can be used to estimate the shape of a plasma subject to magnetic confinement. It was successfully applied, in collaboration with CEA (French nuclear agency) and the University of Nice (JAD Lab.), to data from *Tore Supra* [63]. The procedure is fast because no numerical integration of the underlying PDE is needed, as an explicit basis of solutions to the conjugate Beltrami equation in terms of Bessel functions was found in this case. Generalizing this approach in a more systematic manner to free boundary problems of Bernoulli type, using descent algorithms based on shape-gradient for such approximation-theoretic criteria, is an interesting prospect, still to be pursued.

The piece of work we just mentioned requires defining and studying Hardy spaces of the conjugate-Beltrami equation, which is an interesting topic by itself. For Sobolev-smooth coefficients of exponent greater than 2, this was done in references [4] and [14]. The case of the critical exponent 2 is treated in [34], which apparently provides the first example of well-posedness for the Dirichlet problem in the non-strictly elliptic case: the conductivity may be unbounded or zero on sets of zero capacity and, accordingly, solutions need not be locally bounded.

The 3-D version of step 1 in Section 3.1 is another subject investigated by Apics: to recover a harmonic function (up to a constant) in a ball or a half-space from partial knowledge of its gradient on the boundary. This prototypical inverse problem (*i.e.* inverse to the Cauchy problem for the Laplace equation) often recurs in electromagnetism. At present, Apics is involved with solving instances of this inverse problem arising in two fields, namely medical imaging *e.g.* for electroencephalography (EEG) or magneto-encephalography (MEG), and paleomagnetism (recovery of rocks magnetization) [2], [38], see Section 6.1 . In this connection, we collaborate with two groups of partners: Athena Inria project-team, CHU La Timone, and BESA company on the one hand, Geosciences Lab. at MIT and Cerege CNRS Lab.on the other hand. The question is considerably more difficult than its 2-D counterpart, due mainly to the lack of multiplicative structure for harmonic gradients. Still, considerable progress has been made over the last years using methods of harmonic analysis and operator theory.

The team is further concerned with 3-D generalizations and applications to non-destructive control of step 2 in Section 3.1 . A typical problem is here to localize inhomogeneities or defaults such as cracks, sources or occlusions in a planar or 3-dimensional object, knowing thermal, electrical, or magnetic measurements on the boundary. These defaults can be expressed as a lack of harmonicity of the solution to the associated Dirichlet-Neumann problem, thereby posing an inverse potential problem in order to recover them. In 2-D, finding an optimal discretization of the potential in Sobolev norm amounts to solve a best rational approximation problem, and the question arises as to how the location of the singularities of the approximant (*i.e.* its poles) reflects the location of the singularities of the potential (*i.e.* the defaults we seek). This is a fairly deep issue in approximation theory, to which Apics contributed convergence results for certain classes of fields expressed

as Cauchy integrals over extremal contours for the logarithmic potential [39], [53] [6]. Initial schemes to locate cracks or sources *via* rational approximation on planar domains were obtained this way [56], [43], [46]. It is remarkable that finite inverse source problems in 3-D balls, or more general algebraic surfaces, can be approached using these 2-D techniques upon slicing the domain into planar sections [3], [9]. This bottom line generates a steady research activity within Apics, and again applications are sought to medical imaging and geosciences, see Sections 4.2 , 4.3 and 6.1 .

Conjectures can be raised on the behavior of optimal potential discretization in 3-D, but answering them is an ambitious program still in its infancy.

3.2.2. Systems, transfer and scattering

Participants: Laurent Baratchart, Matthias Caenepeel, Sylvain Chevillard, Sanda Lefteriu, Martine Olivi, Fabien Seyfert.

Through contacts with CNES (French space agency), members of the team became involved in identification and tuning of microwave electromagnetic filters used in space telecommunications, see Section 4.5 . The initial problem was to recover, from band-limited frequency measurements, physical parameters of the device under examination. The latter consists of interconnected dual-mode resonant cavities with negligible loss, hence its scattering matrix is modeled by a 2×2 unitary-valued matrix function on the frequency line, say the imaginary axis to fix ideas. In the bandwidth around the resonant frequency, a modal approximation of the Helmholtz equation in the cavities shows that this matrix is approximately rational, of Mc-Millan degree twice the number of cavities.

This is where system theory comes into play, through the so-called *realization* process mapping a rational transfer function in the frequency domain to a state-space representation of the underlying system of linear differential equations in the time domain. Specifically, realizing the scattering matrix allows one to construct a virtual electrical network, equivalent to the filter, the parameters of which mediate in between the frequency response and the geometric characteristics of the cavities (*i.e.* the tuning parameters).

Hardy spaces provide a framework to transform this ill-posed issue into a series of regularized analytic and meromorphic approximation problems. More precisely, the procedure sketched in Section 3.1 goes as follows:

1. infer from the pointwise boundary data in the bandwidth a stable transfer function (*i.e.* one which is holomorphic in the right half-plane), that may be infinite dimensional (numerically: of high degree). This is done by solving a problem analogous to (P) in Section 3.3.1 , while taking into account prior knowledge on the decay of the response outside the bandwidth, see [13] for details.
2. A stable rational approximation of appropriate degree to the model obtained in the previous step is performed. For this, a descent method on the compact manifold of inner matrices of given size and degree is used, based on an original parametrization of stable transfer functions developed within the team [13].
3. Realizations of this rational approximant are computed. To be useful, they must satisfy certain constraints imposed by the geometry of the device. These constraints typically come from the coupling topology of the equivalent electrical network used to model the filter. This network is composed of resonators, coupled according to some specific graph. This realization step can be recast, under appropriate compatibility conditions [8], as solving a zero-dimensional multivariate polynomial system. To tackle this problem in practice, we use Gröbner basis techniques and continuation methods which team up in the Dedale-HF software (see Section 5.4).

Let us mention that extensions of classical coupling matrix theory to frequency-dependent (reactive) couplings have lately been carried-out [1] for wide-band design applications, although further study is needed to make them computationally effective.

Subsequently Apics started to investigate issues pertaining to design rather than identification. Given the topology of the filter, a basic problem in this connection is to find the optimal response subject to specifications that bear on rejection, transmission and group delay of the scattering parameters. Generalizing the classical approach based on Chebyshev polynomials for single band filters, we recast the problem of multi-band

response synthesis as a generalization of the classical Zolotarev min-max problem for rational functions [29] [11]. Thanks to quasi-convexity, the latter can be solved efficiently using iterative methods relying on linear programming. These were implemented in the software easy-FF (see Section 5.5). Currently, the team is engaged in synthesis of more complex microwave devices like multiplexers and routers, which connect several filters through wave guides. Schur analysis plays an important role here, because scattering matrices of passive systems are of Schur type (*i.e.* contractive in the stability region). The theory originates with the work of I. Schur [77], who devised a recursive test to check for contractivity of a holomorphic function in the disk. The so-called Schur parameters of a function may be viewed as Taylor coefficients for the hyperbolic metric of the disk, and the fact that Schur functions are contractions for that metric lies at the root of Schur's test. Generalizations thereof turn out to be efficient to parametrize solutions to contractive interpolation problems [31]. Dwelling on this, Apics contributed differential parametrizations (atlases of charts) of lossless matrix functions [30][12], [10] which are fundamental to our rational approximation software RARL2 (see Section 5.1). Schur analysis is also instrumental to approach de-embedding issues, and provides one with considerable insight into the so-called matching problem. The latter consists in maximizing the power a multiport can pass to a given load, and for reasons of efficiency it is all-pervasive in microwave and electric network design, *e.g.* of antennas, multiplexers, wifi cards and more. It can be viewed as a rational approximation problem in the hyperbolic metric, and the team presently gets to grips with this hot topic using multipoint contractive interpolation in the framework of the (defense funded) ANR COCORAM, see Sections 6.3.1 and 8.2.1 .

In recent years, our attention was driven by CNES and UPV (Bilbao) to questions about stability of high-frequency amplifiers, see Section 7.2 . Contrary to previously discussed devices, these are *active* components. The response of an amplifier can be linearized around a set of primary current and voltages, and then admittances of the corresponding electrical network can be computed at various frequencies, using the so-called harmonic balance method. The initial goal is to check for stability of the linearized model, so as to ascertain existence of a well-defined working state. The network is composed of lumped electrical elements namely inductors, capacitors, negative *and* positive reactors, transmission lines, and controlled current sources. Our research so far focuses on describing the algebraic structure of admittance functions, so as to set up a function-theoretic framework where the two-steps approach outlined in Section 3.1 can be put to work. The main discovery so far is that the unstable part of each partial transfer function is rational, see Section 6.4 .

3.3. Approximation

Participants: Laurent Baratchart, Sylvain Chevillard, Juliette Leblond, Martine Olivi, Dmitry Ponomarev, Fabien Seyfert.

3.3.1. Best analytic approximation

In dimension 2, the prototypical problem to be solved in step 1 of Section 3.1 may be described as: given a domain $D \subset \mathbb{R}^2$, to recover a holomorphic function from its values on a subset K of the boundary of D. For the discussion it is convenient to normalize D, which can be done by conformal mapping. So, in the simply connected case, we fix D to be the unit disk with boundary unit circle T. We denote by H^p the Hardy space of exponent p, which is the closure of polynomials in $L^p(T)$-norm if $1 \leq p < \infty$ and the space of bounded holomorphic functions in D if $p = \infty$. Functions in H^p have well-defined boundary values in $L^p(T)$, which makes it possible to speak of (traces of) analytic functions on the boundary.

To find an analytic function g in D matching some measured values f approximately on a sub-arc K of T, we formulate a constrained best approximation problem as follows.

> (P) Let $1 \leq p \leq \infty$, K a sub-arc of T, $f \in L^p(K)$, $\psi \in L^p(T \setminus K)$ and $M > 0$; find a function $g \in H^p$ such that $\|g - \psi\|_{L^p(T \setminus K)} \leq M$ and $g - f$ is of minimal norm in $L^p(K)$ under this constraint.

Here ψ is a reference behavior capturing *a priori* assumptions on the behavior of the model off K, while M is some admissible deviation thereof. The value of p reflects the type of stability which is sought and how much one wants to smooth out the data. The choice of L^p classes is suited to handle point-wise measurements.

To fix terminology, we refer to (P) as a *bounded extremal problem*. As shown in [42], [44], [50], the solution to this convex infinite-dimensional optimization problem can be obtained when $p \neq 1$ upon iterating with respect to a Lagrange parameter the solution to spectral equations for appropriate Hankel and Toeplitz operators. These spectral equations involve the solution to the special case $K = T$ of (P), which is a standard extremal problem [66]:

> (P_0) Let $1 \leq p \leq \infty$ and $\varphi \in L^p(T)$; find a function $g \in H^p$ such that $g - \varphi$ is of minimal norm in $L^p(T)$.

The case $p = 1$ is more or less open.

Various modifications of (P) can be set up in order to meet specific needs. For instance when dealing with lossless transfer functions (see Section 4.5), one may want to express the constraint on $T \setminus K$ in a point-wise manner: $|g - \psi| \leq M$ a.e. on $T \setminus K$, see [45]. In this form, the problem comes close to (but still is different from) H^∞ frequency optimization used in control [68], [76]. One can also impose bounds on the real or imaginary part of $g - \psi$ on $T \setminus K$, which is useful when considering Dirichlet-Neuman problems, see [70].

The analog of Problem (P) on an annulus, K being now the outer boundary, can be seen as a means to regularize a classical inverse problem occurring in nondestructive control, namely to recover a harmonic function on the inner boundary from Dirichlet-Neumann data on the outer boundary (see Sections 3.2.1 , 4.2 , 6.1.1 , 6.2). It may serve as a tool to approach Bernoulli type problems, where we are given data on the outer boundary and we *seek the inner boundary*, knowing it is a level curve of the solution.. In this case, the Lagrange parameter indicates how to deform the inner contour in order to improve data fitting. Similar topics are discussed in Sections 3.2.1 and 6.2 for more general equations than the Laplacian, namely isotropic conductivity equations of the form $\mathrm{div}(\sigma \nabla u) = 0$ where σ is no longer constant. Then, the Hardy spaces in Problem (P) are those of a so-called conjugate Beltrami equation: $\overline{\partial} f = \nu \overline{\partial f}$ [69], which are studied for $1 < p < \infty$ in [14], [4], [61] and [34]. Expansions of solutions needed to constructively handle such issues in the specific case of linear fractional conductivities (these occur in plasma shaping) have been expounded in [63].

Though originally considered in dimension 2, Problem (P) carries over naturally to higher dimensions where analytic functions get replaced by gradients of harmonic functions. Namely, given some open set $\Omega \subset \mathbb{R}^n$ and some $\mathbb{R}^n$-valued vector field V on an open subset O of the boundary of Ω, we seek a harmonic function in Ω whose gradient is close to V on O.

When Ω is a ball or a half-space, a substitute for holomorphic Hardy spaces is provided by the Stein-Weiss Hardy spaces of harmonic gradients [80]. Conformal maps are no longer available when $n > 2$, so that Ω can no longer be normalized. More general geometries than spheres and half-spaces have not been much studied so far.

On the ball, the analog of Problem (P) is

> (P_1) Let $1 \leq p \leq \infty$ and $B \subset \mathbb{R}^n$ the unit ball. Fix O an open subset of the unit sphere $S \subset \mathbb{R}^n$. Let further $V \in L^p(O)$ and $W \in L^p(S \setminus O)$ be $\mathbb{R}^n$-valued vector fields. Given $M > 0$, find a harmonic gradient $G \in H^p(B)$ such that $\|G - W\|_{L^p(S \setminus O)} \leq M$ and $G - V$ is of minimal norm in $L^p(O)$ under this constraint.

When $p = 2$, Problem (P_1) was solved in [2] as well as its analog on a shell. The solution extends the one given in [42] for the 2-D case, using a generalization of Toeplitz operators. Thecas of the shell was motivated An important ingredient is a refinement of the Hodge decomposition, that we call the *Hardy-Hodge* decomposition, allowing us to express a $\mathbb{R}^n$-valued vector field in $L^p(S)$, $1 < p < \infty$, as the sum of a vector field in $H^p(B)$, a vector field in $H^p(\mathbb{R}^n \setminus \overline{B})$, and a tangential divergence free vector field on S; the space of such fields is denoted by $D(S)$. If $p = 1$ or $p = \infty$, L^p must be replaced by the real Hardy space or the space of functions with bounded mean oscillation. More generally this decomposition, which is valid on any sufficiently smooth surface (see Section 6.1), seems to play a fundamental role in inverse potential problems. In fact, it was first introduced formally on the plane to describe silent magnetizations supported in $\mathbb{R}^2$ (*i.e.* those generating no field in the upper half space) [38].

Just like solving problem (P) appeals to the solution of problem (P_0), our ability to solve problem (P_1) will depend on the possibility to tackle the special case where $O = S$:

(P_2) Let $1 \leq p \leq \infty$ and $V \in L^p(S)$ be a $\mathbb{R}^n$-valued vector field. Find a harmonic gradient $G \in H^p(B)$ such that $\|G - V\|_{L^p(S)}$ is minimum.

Problem (P_2) is simple when $p = 2$ by virtue of the Hardy Hodge decomposition together with orthogonality of $H^2(B)$ and $H^2(\mathbb{R}^n \setminus \overline{B})$, which is the reason why we were able to solve (P_1) in this case. Other values of p cannot be treated as easily and are currently investigated by Apics, especially the case $p = \infty$ which is of particular interest and presents itself as a 3-D analog to the Nehari problem [75].

Companion to problem (P_2) is problem (P_3) below.

(P_3) Let $1 \leq p \leq \infty$ and $V \in L^p(S)$ be a $\mathbb{R}^n$-valued vector field. Find $G \in H^p(B)$ and $D \in D(S)$ such that $\|G + D - V\|_{L^p(S)}$ is minimum.

Note that (P_2) and (P_3) are identical in 2-D, since no non-constant tangential divergence-free vector field exists on T. It is no longer so in higher dimension, where both (P_2) and (P_3) arise in connection with source recovery in electro/magneto encephalography and paleomagnetism, see Sections 3.2.1 and 4.2 .

3.3.2. Best meromorphic and rational approximation

The techniques set forth in this section are used to solve step 2 in Section 3.2 and instrumental to approach inverse boundary value problems for the Poisson equation $\Delta u = \mu$, where μ is some (unknown) distribution.

3.3.2.1. Scalar meromorphic and rational approximation

We put R_N for the set of rational functions with at most N poles in D. By definition, meromorphic functions in $L^p(T)$ are (traces of) functions in $H^p + R_N$.

A natural generalization of problem (P_0) is:

(P_N) Let $1 \leq p \leq \infty$, $N \geq 0$ an integer, and $f \in L^p(T)$; find a function $g_N \in H^p + R_N$ such that $g_N - f$ is of minimal norm in $L^p(T)$.

Only for $p = \infty$ and f continuous is it known how to solve (P_N) in closed form. The unique solution is given by AAK theory (named after Adamjan, Arov and Krein), which connects the spectral decomposition of Hankel operators with best approximation [75].

The case where $p = 2$ is of special importance for it reduces to rational approximation. Indeed, if we write the Hardy decomposition $f = f^+ + f^-$ where $f^+ \in H^2$ and $f^- \in H^2(\mathbb{C} \setminus \overline{D})$, then $g_N = f^+ + r_N$ where r_N is a best approximant to f^- from R_N in $L^2(T)$. Moreover, r_N has no pole outside D, hence it is a *stable* rational approximant to f^-. However, in contrast to the case where $p = \infty$, this best approximant may *not* be unique.

The former Miaou project (predecessor of Apics) designed a dedicated steepest-descent algorithm for the case $p = 2$ whose convergence to a *local minimum* is guaranteed; until now it seems to be the only procedure meeting this property. This gradient algorithm proceeds recursively with respect to N on a compactification of the parameter space [35]. Although it has proved to be effective in all applications carried out so far (see Sections 4.2 , 4.5), it is still unknown whether the absolute minimum can always be obtained by choosing initial conditions corresponding to *critical points* of lower degree (as is done by the RARL2 software, Section 5.1).

In order to establish global convergence results, Apics has undertaken a deeper study of the number and nature of critical points (local minima, saddle points...), in which tools from differential topology and operator theory team up with classical interpolation theory [47], [49]. Based on this work, uniqueness or asymptotic uniqueness of the approximant was proved for certain classes of functions like transfer functions of relaxation systems (*i.e.* Markov functions) [51] and more generally Cauchy integrals over hyperbolic geodesic arcs [54]. These are the only results of this kind. Research by Apics on this topic remained dormant for a while by reasons of opportunity, but revisiting the work [32] in higher dimension is still a worthy endeavor. Meanwhile,

an analog to AAK theory was carried out for $2 \leq p < \infty$ in [50]. Although not as effective computationally, it was recently used to derive lower bounds [26]. When $1 \leq p < 2$, problem (P_N) is still quite open.

A common feature to the above-mentioned problems is that critical point equations yield non-Hermitian orthogonality relations for the denominator of the approximant. This stresses connections with interpolation, which is a standard way to build approximants, and in many respects best or near-best rational approximation may be regarded as a clever manner to pick interpolation points. This was exploited in [55], [52], and is used in an essential manner to assess the behavior of poles of best approximants to functions with branched singularities, which is of particular interest for inverse source problems (*cf.* Sections 5.6 and 6.1).

In higher dimensions, the analog of Problem (P_N) is best approximation of a vector field by gradients of discrete potentials generated by N point masses. This basic issue is by no means fully understood, and it is an exciting research prospect. It is connected with certain generalizations of Toeplitz or Hankel operators, and with constructive approaches to so-called weak factorizations for real Hardy functions [62].

Besides, certain constrained rational approximation problems, of special interest in identification and design of passive systems, arise when putting additional requirements on the approximant, for instance that it should be smaller than 1 in modulus (*i.e.* a Schur function). In particular, Schur interpolation lately received renewed attention from the team, in connection with matching problems. There, interpolation data are subject to a well-known compatibility condition (positive definiteness of the so-called Pick matrix), and the main difficulty is to put interpolation points on the boundary of D while controlling both the degree and the extremal points of the interpolant. Results obtained by Apics in this direction generalize a variant of contractive interpolation with degree constraint studied in [67], see Section 6.3.1 . We mention that contractive interpolation with nodes approaching the boundary has been a subsidiary research topic by the team in the past, which plays an interesting role in the spectral representation of certain non-stationary stochastic processes [40], [37]. The subject is intimately connected to orthogonal polynomials on the unit circle, and this line of investigation has recently evolved towards an asymptotic study of orthogonal polynomials on planar domains, which is an active area in approximation theory with application to quantum particle systems and Hele-Shaw flows. Section 6.5.1 .

3.3.2.2. Matrix-valued rational approximation

Matrix-valued approximation is necessary to handle systems with several inputs and outputs but it generates additional difficulties as compared to scalar-valued approximation, both theoretically and algorithmically. In the matrix case, the McMillan degree (*i.e.* the degree of a minimal realization in the System-Theoretic sense) generalizes the usual notion of degree for rational functions.

The basic problem that we consider now goes as follows: *let* $\mathcal{F} \in (H^2)^{m\times l}$ *and* n *an integer; find a rational matrix of size* $m \times l$ *without poles in the unit disk and of McMillan degree at most* n *which is nearest possible to* $\mathcal{F}$ *in* $(H^2)^{m\times l}$. Here the L^2 norm of a matrix is the square root of the sum of the squares of the norms of its entries.

The scalar approximation algorithm derived in [35] and mentioned in Section 3.3.2.1 generalizes to the matrix-valued situation [65]. The first difficulty here is to parametrize inner matrices (*i.e.* matrix-valued functions analytic in the unit disk and unitary on the unit circle) of given McMillan degree degree n. Indeed, inner matrices play the role of denominators in fractional representations of transfer matrices (using the so-called Douglas-Shapiro-Shields factorization). The set of inner matrices of given degree is a smooth manifold that allows one to use differential tools as in the scalar case. In practice, one has to produce an atlas of charts (local parametrizations) and to handle changes of charts in the course of the algorithm. Such parametrization can be obtained using interpolation theory and Schur-type algorithms, the parameters of which are vectors or matrices ([30], [10], [12]). Some of these parametrizations are also interesting to compute realizations and achieve filter synthesis ([10] [12]). The rational approximation software “RARL2” developed by the team is described in Section 5.1 .

Difficulties relative to multiple local minima of course arise in the matrix-valued case as well, and deriving criteria that guarantee uniqueness is even more difficult than in the scalar case. The case of rational functions of

degree n or small perturbations thereof (the consistency problem) was solved in [48]. Matrix-valued Markov functions are the only known example beyond this one [33].

Let us stress that RARL2 seems the only algorithm handling rational approximation in the matrix case that demonstrably converges to a local minimum while meeting stability constraints on the approximant.

3.3.3. Behavior of poles of meromorphic approximants

Participant: Laurent Baratchart.

We refer here to the behavior of poles of best meromorphic approximants, in the L^p-sense on a closed curve, to functions f defined as Cauchy integrals of complex measures whose support lies inside the curve. Normalizing the contour to be the unit circle T, we are back to Problem (P_N) in Section 3.3.2.1 ; invariance of the latter under conformal mapping was established in [5]. Research so far has focused on functions whose singular set inside the contour is zero or one-dimensional.

Generally speaking in approximation theory, assessing the behavior of poles of rational approximants is essential to obtain error rates as the degree goes large, and to tackle constructive issues like uniqueness. However, as explained in Section 3.2.1 , Apics considers this issue foremost as a means to extract information on singularities of the solution to a Dirichlet-Neumann problem. The general theme is thus: *how do the singularities of the approximant reflect those of the approximated function?* This approach to inverse problem for the 2-D Laplacian turns out to be attractive when singularities are zero- or one-dimensional (see Section 4.2). It can be used as a computationally cheap initial condition for more precise but much heavier numerical optimizations which often do not even converge unless properly initialized. As regards crack detection or source recovery, this approach boils down to analyzing the behavior of best meromorphic approximants of a function with branch points. For piecewise analytic cracks, or in the case of sources, we were able to prove ([5], [6], [39]), that the poles of the approximants accumulate, when the degree goes large, to some extremal cut of minimum weighted logarithmic capacity connecting the singular points of the crack, or the sources [43]. Moreover, the asymptotic density of the poles turns out to be the Green equilibrium distribution on this cut in D, therefore it charges the singular points if one is able to approximate in sufficiently high degree (this is where the method could fail, because high-order approximation requires rather precise data).

The case of two-dimensional singularities is still an outstanding open problem.

It is remarkable that inverse source problems inside a sphere or an ellipsoid in 3-D can be approached with such 2-D techniques, as applied to planar sections (see Section 6.1). The technique is implemented in the software FindSources3D, see Section 5.6 .

3.3.4. Miscellaneous

Participant: Sylvain Chevillard.

Sylvain Chevillard, joined team in November 2010. His coming resulted in Apics hosting a research activity in certified computing, centered on the software *Sollya* of which S. Chevillard is a co-author, see Section 5.7 . On the one hand, Sollya is an Inria software which still requires some tuning to a growing community of users. On the other hand, approximation-theoretic methods at work in Sollya are potentially useful for certified solutions to constrained analytic problems described in Section 3.3.1 . However, developing Sollya is not a long-term objective of Apics.

ASPI Project-Team

3. Research Program

3.1. Interacting Monte Carlo methods and particle approximation of Feynman–Kac distributions

Monte Carlo methods are numerical methods that are widely used in situations where (i) a stochastic (usually Markovian) model is given for some underlying process, and (ii) some quantity of interest should be evaluated, that can be expressed in terms of the expected value of a functional of the process trajectory, which includes as an important special case the probability that a given event has occurred. Numerous examples can be found, e.g. in financial engineering (pricing of options and derivative securities) [36], in performance evaluation of communication networks (probability of buffer overflow), in statistics of hidden Markov models (state estimation, evaluation of contrast and score functions), etc. Very often in practice, no analytical expression is available for the quantity of interest, but it is possible to simulate trajectories of the underlying process. The idea behind Monte Carlo methods is to generate independent trajectories of this process or of an alternate instrumental process, and to build an approximation (estimator) of the quantity of interest in terms of the weighted empirical probability distribution associated with the resulting independent sample. By the law of large numbers, the above estimator converges as the size N of the sample goes to infinity, with rate $1/\sqrt{N}$ and the asymptotic variance can be estimated using an appropriate central limit theorem. To reduce the variance of the estimator, many variance reduction techniques have been proposed. Still, running independent Monte Carlo simulations can lead to very poor results, because trajectories are generated *blindly*, and only afterwards are the corresponding weights evaluated. Some of the weights can happen to be negligible, in which case the corresponding trajectories are not going to contribute to the estimator, i.e. computing power has been wasted.

A recent and major breakthrough, has been the introduction of interacting Monte Carlo methods, also known as sequential Monte Carlo (SMC) methods, in which a whole (possibly weighted) sample, called *system of particles*, is propagated in time, where the particles

- *explore* the state space under the effect of a *mutation* mechanism which mimics the evolution of the underlying process,
- and are *replicated* or *terminated*, under the effect of a *selection* mechanism which automatically concentrates the particles, i.e. the available computing power, into regions of interest of the state space.

In full generality, the underlying process is a discrete–time Markov chain, whose state space can be

> finite, continuous, hybrid (continuous / discrete), graphical, constrained, time varying, pathwise, etc.,

the only condition being that it can easily be *simulated*.

In the special case of particle filtering, originally developed within the tracking community, the algorithms yield a numerical approximation of the optimal Bayesian filter, i.e. of the conditional probability distribution of the hidden state given the past observations, as a (possibly weighted) empirical probability distribution of the system of particles. In its simplest version, introduced in several different scientific communities under the name of *bootstrap filter* [38], *Monte Carlo filter* [43] or *condensation* (conditional density propagation) algorithm [40], and which historically has been the first algorithm to include a redistribution step, the selection mechanism is governed by the likelihood function: at each time step, a particle is more likely to survive and to replicate at the next generation if it is consistent with the current observation. The algorithms also provide as a by–product a numerical approximation of the likelihood function, and of many other contrast functions for parameter estimation in hidden Markov models, such as the prediction error or the conditional least–squares criterion.

Particle methods are currently being used in many scientific and engineering areas

> positioning, navigation, and tracking [39], [33], visual tracking [40], mobile robotics [34], [55], ubiquitous computing and ambient intelligence, sensor networks, risk evaluation and simulation of rare events [37], genetics, molecular simulation [35], etc.

Other examples of the many applications of particle filtering can be found in the contributed volume [22] and in the special issue of *IEEE Transactions on Signal Processing* devoted to *Monte Carlo Methods for Statistical Signal Processing* in February 2002, where the tutorial paper [23] can be found, and in the textbook [52] devoted to applications in target tracking. Applications of sequential Monte Carlo methods to other areas, beyond signal and image processing, e.g. to genetics, can be found in [51]. A recent overview can also be found in [25].

Particle methods are very easy to implement, since it is sufficient in principle to simulate independent trajectories of the underlying process. The whole problematic is multidisciplinary, not only because of the already mentioned diversity of the scientific and engineering areas in which particle methods are used, but also because of the diversity of the scientific communities which have contributed to establish the foundations of the field

> target tracking, interacting particle systems, empirical processes, genetic algorithms (GA), hidden Markov models and nonlinear filtering, Bayesian statistics, Markov chain Monte Carlo (MCMC) methods.

These algorithms can be interpreted as numerical approximation schemes for Feynman–Kac distributions, a pathwise generalization of Gibbs–Boltzmann distributions, in terms of the weighted empirical probability distribution associated with a system of particles. This abstract point of view [31], [29], has proved to be extremely fruitful in providing a very general framework to the design and analysis of numerical approximation schemes, based on systems of branching and / or interacting particles, for nonlinear dynamical systems with values in the space of probability distributions, associated with Feynman–Kac distributions. Many asymptotic results have been proved as the number N of particles (sample size) goes to infinity, using techniques coming from applied probability (interacting particle systems, empirical processes [56]), see e.g. the survey article [31] or the textbooks [29], [28], and references therein

> convergence in ^{p}, convergence as empirical processes indexed by classes of functions, uniform convergence in time, see also [48], [49], central limit theorem, see also [45], propagation of chaos, large deviations principle, etc.

The objective here is to systematically study the impact of the many algorithmic variants on the convergence results.

3.2. Statistics of HMM

Hidden Markov models (HMM) form a special case of partially observed stochastic dynamical systems, in which the state of a Markov process (in discrete or continuous time, with finite or continuous state space) should be estimated from noisy observations. The conditional probability distribution of the hidden state given past observations is a well–known example of a normalized (nonlinear) Feynman–Kac distribution, see 3.1 . These models are very flexible, because of the introduction of latent variables (non observed) which allows to model complex time dependent structures, to take constraints into account, etc. In addition, the underlying Markovian structure makes it possible to use numerical algorithms (particle filtering, Markov chain Monte Carlo methods (MCMC), etc.) which are computationally intensive but whose complexity is rather small. Hidden Markov models are widely used in various applied areas, such as speech recognition, alignment of biological sequences, tracking in complex environment, modeling and control of networks, digital communications, etc.

Beyond the recursive estimation of a hidden state from noisy observations, the problem arises of statistical inference of HMM with general state space [26], including estimation of model parameters, early monitoring and diagnosis of small changes in model parameters, etc.

Large time asymptotics A fruitful approach is the asymptotic study, when the observation time increases to infinity, of an extended Markov chain, whose state includes (i) the hidden state, (ii) the observation, (iii) the prediction filter (i.e. the conditional probability distribution of the hidden state given observations at all previous time instants), and possibly (iv) the derivative of the prediction filter with respect to the parameter. Indeed, it is easy to express the log–likelihood function, the conditional least–squares criterion, and many other clasical contrast processes, as well as their derivatives with respect to the parameter, as additive functionals of the extended Markov chain.

The following general approach has been proposed

- first, prove an exponential stability property (i.e. an exponential forgetting property of the initial condition) of the prediction filter and its derivative, for a misspecified model,
- from this, deduce a geometric ergodicity property and the existence of a unique invariant probability distribution for the extended Markov chain, hence a law of large numbers and a central limit theorem for a large class of contrast processes and their derivatives, and a local asymptotic normality property,
- finally, obtain the consistency (i.e. the convergence to the set of minima of the associated contrast function), and the asymptotic normality of a large class of minimum contrast estimators.

This programme has been completed in the case of a finite state space [7], and has been generalized [32] under an uniform minoration assumption for the Markov transition kernel, which typically does only hold when the state space is compact. Clearly, the whole approach relies on the existence of an exponential stability property of the prediction filter, and the main challenge currently is to get rid of this uniform minoration assumption for the Markov transition kernel [30], [49], so as to be able to consider more interesting situations, where the state space is noncompact.

Small noise asymptotics Another asymptotic approach can also be used, where it is rather easy to obtain interesting explicit results, in terms close to the language of nonlinear deterministic control theory [44]. Taking the simple example where the hidden state is the solution to an ordinary differential equation, or a nonlinear state model, and where the observations are subject to additive Gaussian white noise, this approach consists in assuming that covariances matrices of the state noise and of the observation noise go simultaneously to zero. If it is reasonable in many applications to consider that noise covariances are small, this asymptotic approach is less natural than the large time asymptotics, where it is enough (provided a suitable ergodicity assumption holds) to accumulate observations and to see the expected limit laws (law of large numbers, central limit theorem, etc.). In opposition, the expressions obtained in the limit (Kullback–Leibler divergence, Fisher information matrix, asymptotic covariance matrix, etc.) take here a much more explicit form than in the large time asymptotics.

The following results have been obtained using this approach

- the consistency of the maximum likelihood estimator (i.e. the convergence to the set M of global minima of the Kullback–Leibler divergence), has been obtained using large deviations techniques, with an analytical approach [41],
- if the abovementioned set M does not reduce to the true parameter value, i.e. if the model is not identifiable, it is still possible to describe precisely the asymptotic behavior of the estimators [42]: in the simple case where the state equation is a noise–free ordinary differential equation and using a Bayesian framework, it has been shown that (i) if the rank r of the Fisher information matrix I is constant in a neighborhood of the set M, then this set is a differentiable submanifold of codimension r, (ii) the posterior probability distribution of the parameter converges to a random probability distribution in the limit, supported by the manifold M, absolutely continuous w.r.t. the Lebesgue measure on M, with an explicit expression for the density, and (iii) the posterior probability distribution of the suitably normalized difference between the parameter and its projection on the manifold M, converges to a mixture of Gaussian probability distributions on the normal spaces to the manifold M, which generalized the usual asymptotic normality property,

- it has been shown [50] that (i) the parameter dependent probability distributions of the observations are locally asymptotically normal (LAN) [47], from which the asymptotic normality of the maximum likelihood estimator follows, with an explicit expression for the asymptotic covariance matrix, i.e. for the Fisher information matrix I, in terms of the Kalman filter associated with the linear tangent linear Gaussian model, and (ii) the score function (i.e. the derivative of the log–likelihood function w.r.t. the parameter), evaluated at the true value of the parameter and suitably normalized, converges to a Gaussian r.v. with zero mean and covariance matrix I.

3.3. Multilevel splitting for rare event simulation

See 4.2 , and 5.1 , 5.2 , and 5.3 .

The estimation of the small probability of a rare but critical event, is a crucial issue in industrial areas such as

> nuclear power plants, food industry, telecommunication networks, finance and insurance industry, air traffic management, etc.

In such complex systems, analytical methods cannot be used, and naive Monte Carlo methods are clearly unefficient to estimate accurately very small probabilities. Besides importance sampling, an alternate widespread technique consists in multilevel splitting [46], where trajectories going towards the critical set are given offsprings, thus increasing the number of trajectories that eventually reach the critical set. As shown in [5], the Feynman–Kac formalism of 3.1 is well suited for the design and analysis of splitting algorithms for rare event simulation.

Propagation of uncertainty Multilevel splitting can be used in static situations. Here, the objective is to learn the probability distribution of an output random variable $Y = F(X)$, where the function F is only defined pointwise for instance by a computer programme, and where the probability distribution of the input random variable X is known and easy to simulate from. More specifically, the objective could be to compute the probability of the output random variable exceeding a threshold, or more generally to evaluate the cumulative distribution function of the output random variable for different output values. This problem is characterized by the lack of an analytical expression for the function, the computational cost of a single pointwise evaluation of the function, which means that the number of calls to the function should be limited as much as possible, and finally the complexity and / or unavailability of the source code of the computer programme, which makes any modification very difficult or even impossible, for instance to change the model as in importance sampling methods.

The key issue is to learn as fast as possible regions of the input space which contribute most to the computation of the target quantity. The proposed splitting methods consists in (i) introducing a sequence of intermediate regions in the input space, implicitly defined by exceeding an increasing sequence of thresholds or levels, (ii) counting the fraction of samples that reach a level given that the previous level has been reached already, and (iii) improving the diversity of the selected samples, usually using an artificial Markovian dynamics. In this way, the algorithm learns

- the transition probability between successive levels, hence the probability of reaching each intermediate level,
- and the probability distribution of the input random variable, conditionned on the output variable reaching each intermediate level.

A further remark, is that this conditional probability distribution is precisely the optimal (zero variance) importance distribution needed to compute the probability of reaching the considered intermediate level.

Rare event simulation To be specific, consider a complex dynamical system modelled as a Markov process, whose state can possibly contain continuous components and finite components (mode, regime, etc.), and the objective is to compute the probability, hopefully very small, that a critical region of the state space is reached by the Markov process before a final time T, which can be deterministic and fixed, or random (for instance the time of return to a recurrent set, corresponding to a nominal behaviour).

The proposed splitting method consists in (i) introducing a decreasing sequence of intermediate, more and more critical, regions in the state space, (ii) counting the fraction of trajectories that reach an intermediate region before time T, given that the previous intermediate region has been reached before time T, and (iii) regenerating the population at each stage, through redistribution. In addition to the non–intrusive behaviour of the method, the splitting methods make it possible to learn the probability distribution of typical critical trajectories, which reach the critical region before final time T, an important feature that methods based on importance sampling usually miss. Many variants have been proposed, whether

- the branching rate (number of offsprings allocated to a successful trajectory) is fixed, which allows for depth–first exploration of the branching tree, but raises the issue of controlling the population size,
- the population size is fixed, which requires a breadth–first exploration of the branching tree, with random (multinomial) or deterministic allocation of offsprings, etc.

Just as in the static case, the algorithm learns

- the transition probability between successive levels, hence the probability of reaching each intermediate level,
- and the entrance probability distribution of the Markov process in each intermediate region.

Contributions have been given to

- minimizing the asymptotic variance, obtained through a central limit theorem, with respect to the shape of the intermediate regions (selection of the importance function), to the thresholds (levels), to the population size, etc.
- controlling the probability of extinction (when not even one trajectory reaches the next intermediate level),
- designing and studying variants suited for hybrid state space (resampling per mode, marginalization, mode aggregation),

and in the static case, to

- minimizing the asymptotic variance, obtained through a central limit theorem, with respect to intermediate levels, to the Metropolis kernel introduced in the mutation step, etc.

A related issue is global optimization. Indeed, the difficult problem of finding the set M of global minima of a real–valued function V can be replaced by the apparently simpler problem of sampling a population from a probability distribution depending on a small parameter, and asymptotically supported by the set M as the small parameter goes to zero. The usual approach here is to use the cross–entropy method [53], [27], which relies on learning the optimal importance distribution within a prescribed parametric family. On the other hand, multilevel splitting methods could provide an alternate nonparametric approach to this problem.

3.4. Nearest neighbor estimates

This additional topic was not present in the initial list of objectives, and has emerged only recently.

In pattern recognition and statistical learning, also known as machine learning, nearest neighbor (NN) algorithms are amongst the simplest but also very powerful algorithms available. Basically, given a training set of data, i.e. an N–sample of i.i.d. object–feature pairs, with real–valued features, the question is how to generalize, that is how to guess the feature associated with any new object. To achieve this, one chooses some integer k smaller than N, and takes the mean–value of the k features associated with the k objects that are nearest to the new object, for some given metric.

In general, there is no way to guess exactly the value of the feature associated with the new object, and the minimal error that can be done is that of the Bayes estimator, which cannot be computed by lack of knowledge of the distribution of the object–feature pair, but the Bayes estimator can be useful to characterize the strength of the method. So the best that can be expected is that the NN estimator converges, say when the sample size N grows, to the Bayes estimator. This is what has been proved in great generality by Stone [54] for the mean square convergence, provided that the object is a finite–dimensional random variable, the feature is a square–integrable random variable, and the ratio k/N goes to 0. Nearest neighbor estimator is not the only local averaging estimator with this property, but it is arguably the simplest.

The asymptotic behavior when the sample size grows is well understood in finite dimension, but the situation is radically different in general infinite dimensional spaces, when the objects to be classified are functions, images, etc.

Nearest neighbor classification in infinite dimension In finite dimension, the k–nearest neighbor classifier is universally consistent, i.e. its probability of error converges to the Bayes risk as N goes to infinity, whatever the joint probability distribution of the pair, provided that the ratio k/N goes to zero. Unfortunately, this result is no longer valid in general metric spaces, and the objective is to find out reasonable sufficient conditions for the weak consistency to hold. Even in finite dimension, there are exotic distances such that the nearest neighbor does not even get closer (in the sense of the distance) to the point of interest, and the state space needs to be complete for the metric, which is the first condition. Some regularity on the regression function is required next. Clearly, continuity is too strong because it is not required in finite dimension, and a weaker form of regularity is assumed. The following consistency result has been obtained: if the metric space is separable and if some Besicovich condition holds, then the nearest neighbor classifier is weakly consistent. Note that the Besicovich condition is always fulfilled in finite dimensional vector spaces (this result is called the Besicovich theorem), and that a counterexample [3] can be given in an infinite dimensional space with a Gaussian measure (in this case, the nearest neighbor classifier is clearly nonconsistent). Finally, a simple example has been found which verifies the Besicovich condition with a noncontinuous regression function.

Rates of convergence of the functional k–nearest neighbor estimator Motivated by a broad range of potential applications, such as regression on curves, rates of convergence of the k–nearest neighbor estimator of the regression function, based on N independent copies of the object–feature pair, have been investigated when the object is in a suitable ball in some functional space. Using compact embedding theory, explicit and general finite sample bounds can be obtained for the expected squared difference between the k–nearest neighbor estimator and the Bayes regression function, in a very general setting. The results have also been particularized to classical function spaces such as Sobolev spaces, Besov spaces and reproducing kernel Hilbert spaces. The rates obtained are genuine nonparametric convergence rates, and up to our knowledge the first of their kind for k–nearest neighbor regression.

This emerging topic has produced several theoretical advances [1], [2] in collaboration with Gérard Biau (université Pierre et Marie Curie, ENS Paris and EPI CLASSIC, Inria Paris—Rocquencourt), and a possible target application domain has been identified in the statistical analysis of recommendation systems, that would be a source of interesting problems.

BACCHUS Team

3. Research Program

3.1. Numerical schemes for fluid mechanics

Participants: Luca Arpaia, Héloïse Beaugendre, Pietro Marco Congedo, Cécile Dobrzynski, Andrea Filippini, Maria Kazolea, Luc Mieussens, Mario Ricchiuto, Maria Giovanna Rodio.

A large number of engineering problems involve fluid mechanics. They may involve the coupling of one or more physical models. An example is provided by aeroelastic problems, which have been studied in details by other Inria teams. Another example is given by flows in pipelines where the fluid (a mixture of air–water–gas) does not have well-known physical properties, and there are even more exotic situations. In some occasions, one needs specific numerical tools to take into account *e.g.* a fluids' exotic equation of state, or a the influence of small flow scales in a macro-/meso-scopic flow model, etc. Efficient schemes are needed in unsteady flows where the amount of required computational resources becomes huge. Another situation where specific tools are needed is when one is interested in very specific physical quantities, such as *e.g.* the lift and drag of an airfoil, or the boundary of the area flooded by a Tsunami.

In these situations, commercial tools can only provide a crude answer. These codes, while allowing users to simulate a lot of different flow types, and "always" providing an answer, often give results of poor quality. This is mainly due to their general purpose character, and on the fact that the numerical technology implemented in these codes is not the most recent. To give a few examples, consider the noise generated by wake vortices in supersonic flows (external aerodynamics/aeroacoustics), or the direct simulation of a 3D compressible mixing layer in a complex geometry (as in combustion chambers). Up to our knowledge, due to the very different temporal and physical scales that need to be captured, a direct simulation of these phenomena is not in the reach of the most recent technologies because the numerical resources required are currently unavailable. *We need to invent specific algorithms for this purpose.*

Our goal is to develop more accurate and more efficient schemes that can adapt to modern computer architectures, and allow the efficient simulation of complex real life flows.

We develop a class of numerical schemes, known in literature as Residual Distribution schemes, *specifically tailored to unstructured and hybrid meshes*. They have the most possible compact stencil that is compatible with the expected order of accuracy. This *accuracy is at least of second order, and it can go up to any order of accuracy, even though fourth order is considered for practical applications.* Since the stencil is compact, the implementation on parallel machines becomes simple. These schemes are very flexible in nature, which is so far one of the most important advantage over other techniques. This feature has allowed us to adapt the schemes to the requirements of different physical situations (*e.g.* different formulations allow either en efficient explicit time advancement for problems involving small time-scales, or a fully implicit space-time variant which is unconditionally stable and allows to handle stiff problems where only the large time scales are relevant). This flexibility has also enabled to devise a variant using the same data structure of the popular Discontinuous Galerkin schemes, which are also part of our scientific focus.

The compactness of the second order version of the schemes enables us to use efficiently the high performance parallel linear algebra tools developed by the team. However, the high order versions of these schemes, which are under development, require modifications to these tools taking into account the nature of the data structure used to reach higher orders of accuracy. This leads to new scientific problems at the border between numerical analysis and computer science. In parallel to these fundamental aspects, we also work on adapting more classical numerical tools to complex physical problems such as those encountered in interface flows, turbulent or multiphase flows, geophysical flows, and material science. A particular attention has been devoted to the implementation of complex thermodynamic models permitting to simulate several classes of fluids and to take into account real-gas effects and some exotic phenomenon, such as rarefaction shock waves.

Within these applications, a strong effort has been made in developing more predictive tools for both multiphase compressible flows and non-hydrostatic free surface flows.

Concerning multiphase flows, several advancements have been performed, i.e. considering a more complete systems of equations including viscosity, working on the thermodynamic modelling of complex fluids, and developing stochastic methods for uncertainty quantification in compressible flows. Concerning depth averaged free surface flow modelling, on one hand we have shown the advantages of the use of the compact schemes we develop for hydrostatic shallow water models. On the other, we have shown ho to extend our approach to non-hydrostatic Boussinesq modelling, including wave dispersion, and wave breaking effects.
We expect to be able to demonstrate the potential of our developments on applications ranging from the the reproduction of the complex multidimensional interactions between tidal waves and estuaries, to the unsteady aerodynamics and aeroacoustics associated to both external and internal compressible flows, and the behaviour of complex materials. This will be achieved by means of a multi-disciplinary effort involving our research on residual discretizations schemes, the parallel advances in algebraic solvers and partitioners, and the strong interactions with specialists in computer science, scientific computing, physics, mechanics, and mathematical modeling.
Concerning the software platforms, our research in numerical algorithms has led to the development of the `RealfluiDS` platform which is described in section 4.3 , and to the `SLOWS` (Shallow-water fLOWS) code for free surface flows, described in sections 4.10 . Simultaneously, we have contributed to the advancement of the new, object oriented, parallel finite elements library `AeroSol`, described in section 4.1 , which is destined to replace the existing codes and become the team's CFD kernel. Concerning uncertainty quantification and robust optimization, we are developing the platform `RobUQ`.

New software developments are under way in the field of complex materials modeling and multiphase flows with heat and mass transfer. Concerning the materials modelling, these developments are performed in the code in the solver `COCA` (CodeOxydationCompositesAutocicatrisants) for the simulation of the self-healing process in composite materials. These developments will be described in section 4.2 . Concerning the numerical simulation of multiphase flows, we have developed the code `sDEM`, which is one of rare code, permitting to simulate metastable states with a complex thermodynamics and considering uncertainty quantification techniques.
Funding and external collaborations. This work is supported by several sources including the last of the `ADDECCO` ERC grant, the FP7 STORM, the ANR `UFO` and the PIA project `TANDEM`. Important contributions to these activities are given by our external collaborators, and in particular R. Abgrall (Universität Zürich), P. Bonneton (UMR EPOC Bordeaux), G. Vignoles (LCTS lab Bordeaux), and D. De Santis (via the associated team `AQUARIUS`).

3.2. Numerical schemes for Uncertainty quantification and robust optimization

Participants: Pietro Marco Congedo, Francesca Fusi, Gianluca Geraci, Mario Ricchiuto, Maria Giovanna Rodio, Kunkun Tang.

Another topic of interest is the quantification of uncertainties in non linear problems. In many applications, the physical model is not known accurately. The typical example is that of turbulence models in aeronautics. These models all depend on a number of parameters which can radically change the output of the simulation. Being impossible to lump the large number of temporal and spatial scales of a turbulent flow in a few model parameters, these values are often calibrated to quantitatively reproduce a certain range of effects observed experimentally. A similar situation is encountered in many applications such as real gas or multiphase flows, where the equation of state form suffer from uncertainties, and free surface flows with sediment transport, where often both the hydrodynamic model and the sediment transport model depend on several parameters, and my have more than one formal expression.

This type of uncertainty, called *epistemic*, is associated with a lack of knowledge and could be reduced by further experiments and investigation. Instead, another type of uncertainty, called *aleatory*, is related to the intrinsec aleatory quality of a physical measure and can not be reduced. The dependency of the numerical simulation from these uncertainties can be studied by propagation of chaos techniques such as those developed during the recent years via polynomial chaos techniques. Different implementations exists, depending whether the method is intrusive or not. The accuracy of these methods is still a matter of research, as well how they can handle an as large as possible number of uncertainties or their versatility with respect to the structure of the random variable pdfs.

Our objective is to develop some non-intrusive and semi-intrusive methods, trying to define an unified framework for obtained a reliable and accurate numerical solution at a moderate computational cost. This work have produced a large number of publications on peer-reviewed journal. Concerning the class of intrusive methods, we are developing an unified scheme in the coupled physical/stochastic space based on a multi-resolution framework. Here, the idea is to build a framework for being capable to refine a discontinuity in both stochastic and deterministic mesh. We are extending this class of methods to complex models in CFD, such as in multiphase flows. Concerning the non-intrusive methods, we are working on several methods for treating the following problems : handling a large number of uncertainties, treating high-order statistical decomposition (variance, skewness and kurtosis), and solving efficiently inverse problems.

We have used these methods to several ends: either to have highly accurate quantitative reconstruction of a simulation output's variation over a complex space of parameter variations to study a given model (uncertainty propagation), or as a means of comparing different model's variability to certain parameters thus assessing their robustness (model robustness), or as a tool to compare different numerical implementation (schemes and codes) of a similar model to assess simultaneously the robustness of the numerics and the universality of the trends of the statistics and of the sensitivity measures (robust cross-validation). Moreover, we rebuild statistically some input parameters relying on some experimental measures of the output, thus solving an inverse problem.

The developed methods and tools have been applied to several applications of interest : real-gas effects, multi-phase flows, cavitation, aerospace applications and geophysical flows.

Concerning robust optimization, we focus on problems with high dimensional representation of stochastic inputs, that can be computationally prohibitive. In fact, for a robust design, statistics of the fitness functions are also important, then uncertainty quantification (UQ) becomes the predom- inant issue to handle if a large number of uncertainties is taken into account. Several methods are proposed in literature to consider high dimension stochastic problem but their accuracy on realistic problems where highly non-linear effects could exist is not proven at all. We developed several efficient global strategies for robust optimization: the first class of method is based on the extension of simplex stochastic collocation to the optimization space, the second one consists in hybrid strategies using ANOVA decomposition.

These developments and computations are performed in the platform `RobUQ`, which includes the most part of methods developed in the Team.
Funding and external collaborations. This part of our activities is supported by the ANR-MN project `UFO`, and the associated team `AQUARIUS`. It benefits from the collaborations with external members, and in particular R. Abgrall (Universität Zürich), and of the members of the associated team.

3.3. Meshes and scalable discrete data structures

Participants: Luca Arpaia, Cécile Dobrzynski, Algiane Froehly, Cédric Lachat, François Pellegrini, Mario Ricchiuto.

3.3.1. Dynamic mesh adaptation and partitioning

Many simulations which model the evolution of a given phenomenon along with time (turbulence and unsteady flows, for instance) need to re-mesh some portions of the problem graph in order to capture more accurately the properties of the phenomenon in areas of interest. This re-meshing is performed according to criteria which

are closely linked to the undergoing computation and can involve large mesh modifications: while elements are created in critical areas, some may be merged in areas where the phenomenon is no longer critical. To alleviate the cost of this re-meshing phase, we have started looking into time dependent continuous mesh deformation techniques. These may allow some degree of adaptation between two re-meshing phases, which in theory may be less frequent, and more local.

When working in parallel, re-meshing introduces additional problems. In particular, splitting an element which is located on the frontier between several processors is not an easy task, because deciding when splitting some element, and defining the direction along which to split it so as to preserve numerical stability most, require shared knowledge which is not available in distributed memory architectures. Ad-hoc data structures and algorithms have to be devised so as to achieve these goals without resorting to extra communication and synchronization which would impact the running speed of the simulation.

Most of the works on parallel mesh adaptation attempt to parallelize in some way all the mesh operations: edge swap, edge split, point insertion, etc. It implies deep modifications in the (re)mesher and often leads to bad performance in term of CPU time. An other work [54] proposes to base the parallel re-meshing on existing mesher and load balancing to be able to modify the elements located on the frontier between several processors.

In addition, the preservation of load balance in the re-meshed simulation requires dynamic redistribution of mesh data across processing elements. Several dynamic repartitioning methods have been proposed in the literature [55], [53], which rely on diffusion-like algorithms and the solving of flow problems to minimize the amount of data to be exchanged between processors. However, integrating such algorithms into a global framework for handling adaptive meshes in parallel has yet to be done.

The path that we are following bases on the decomposition of the areas to remesh into balls that can be processed concurrently, each by a sequential remesher. It requires to devise scalable algorithms for building such boules, scheduling them on as many processors as possible, reconstructing the remeshed mesh and redistributing its data.
Funding and external collaborations. Most of this research has started within the context of the PhD of Cédric Lachat, funded by a CORDI grant of EPI `PUMAS` and was continued thanks to a funding by ADT grant `El Gaucho` that completed this year. The work on adaptation by continuous deformation has started with the PhD of L. Arpaia and benefits of the funding of the PIA project `TANDEM`.

3.3.2. Graph partitioning and static mapping

Unlike their predecessors of two decades ago, today's very large parallel architectures can no longer implement a uniform memory model. They are based on a hierarchical structure, in which cores are assembled into chips, chips are assembled into boards, boards are assembled into cabinets and cabinets are interconnected through high speed, low latency communication networks. On these systems, communication is non uniform: communicating with cores located on the same chip is cheaper than with cores on other boards, and much cheaper than with cores located in other cabinets. The advent of these massively parallel, non uniform machines impacts the design of the software to be executed on them, both for applications and for service tools. It is in particular the case for the software whose task is to balance workload across the cores of these architectures.

A common method for task allocation is to use graph partitioning tools. The elementary computations to perform are represented by vertices and their dependencies by edges linking two vertices that need to share some piece of data. Finding good solutions to the workload distribution problem amounts to computing partitions with small vertex or edge cuts and that balance evenly the weights of the graph parts. Yet, computing efficient partitions for non uniform architectures requires to take into account the topology of the target architecture. When processes are assumed to coexist simultaneously for all the duration of the program, this generalized optimization problem is called mapping. In this problem, the communication cost function to minimize incorporates architecture-dependent, locality improving terms, such as the dilation of each edge (that is, by how much it is "stretched" across the graph representing the target architecture), which is sometimes

also expressed as some “hop metric”. A mapping is called static if it is computed prior to the execution of the program and is never modified at run-time.

The sequential Scotch tool being developed within the BACCHUS team (see Section 4.9) was able to perform static mapping since its first version, in 1994, but this feature was not widely known nor used by the community. With the increasing need to map very large problem graphs onto very large and strongly non uniform parallel machines, there is an increasing demand for parallel static mapping tools. Since, in the context of dynamic repartitioning, parallel mapping software will have to run on their target architectures, parallel mapping and remapping algorithms suitable for efficient execution on such heterogeneous architectures have to be investigated. This leads to solve three interwoven challenges:

- scalability: such algorithms must be able to map graphs of more than a billion vertices onto target architectures comprising millions of cores;
- heterogeneity: not only do these algorithms must take into account the topology of the target architecture they map graphs onto, but they also have themselves to run efficiently on these very architectures;
- asynchronicity: most parallel partitioning algorithms use collective communication primitives, that is, some form of heavy synchronization. With the advent of machines having several millions of cores, and in spite of the continuous improvement of communication subsystems, the demand for more asynchronicity in parallel algorithms is likely to increase.

This year, our work mostly concerned the tighter integration of Scotch with PaMPA. In particular, the routines for partitioning with fixed vertices, which are mandatory in PaMPA to balance remeshing workload across processing elements that already contain some mesh data, have been redesigned almost from scratch.

BIPOP Project-Team

3. Research Program

3.1. Dynamic non-regular systems

mechanical systems, impacts, unilateral constraints, complementarity, modeling, analysis, simulation, control, convex analysis

Dynamical systems (we limit ourselves to finite-dimensional ones) are said to be *non-regular* whenever some nonsmoothness of the state arises. This nonsmoothness may have various roots: for example some outer impulse, entailing so-called *differential equations with measure*. An important class of such systems can be described by the complementarity system

$$\begin{cases} \dot{x} = f(x,u,\lambda)\,, \\ 0 \le y \perp \lambda \ge 0\,, \\ g(y,\lambda,x,u,t) = 0\,, \\ \text{re-initialization law of the state } x(\cdot)\ , \end{cases} \tag{3}$$

where $\perp$ denotes orthogonality; u is a control input. Now (1) can be viewed from different angles.

- Hybrid systems: it is in fact natural to consider that (1) corresponds to different models, depending whether $y_i = 0$ or $y_i > 0$ (y_i being a component of the vector y). In some cases, passing from one mode to the other implies a jump in the state x; then the continuous dynamics in (1) may contain distributions.
- Differential inclusions: $0 \le y \perp \lambda \ge 0$ is equivalent to $-\lambda \in \mathrm{N}_K(y)$, where K is the nonnegative orthant and $\mathrm{N}_K(y)$ denotes the normal cone to K at y. Then it is not difficult to reformulate (1) as a differential inclusion.
- Dynamic variational inequalities: such a formalism reads as $\langle \dot{x}(t) + F(x(t),t), v - x(t)\rangle \ge 0$ for all $v \in K$ and $x(t) \in K$, where K is a nonempty closed convex set. When K is a polyhedron, then this can also be written as a complementarity system as in (1).

Thus, the 2nd and 3rd lines in (1) define the modes of the hybrid systems, as well as the conditions under which transitions occur from one mode to another. The 4th line defines how transitions are performed by the state x. There are several other formalisms which are quite related to complementarity. A tutorial-survey paper has been published [5], whose aim is to introduce the dynamics of complementarity systems and the main available results in the fields of mathematical analysis, analysis for control (controllability, observability, stability), and feedback control.

3.2. Nonsmooth optimization

optimization, numerical algorithm, convexity, Lagrangian relaxation, combinatorial optimization.

Here we are dealing with the minimization of a function f (say over the whole space R^n), whose derivatives are discontinuous. A typical situation is when f comes from dualization, if the primal problem is not strictly convex – for example a large-scale linear program – or even nonconvex – for example a combinatorial optimization problem. Also important is the case of spectral functions, where $f(x) = F(\lambda(A(x)))$, A being a symmetric matrix and λ its spectrum.

For these types of problems, we are mainly interested in developing efficient resolution algorithms. Our basic tool is bundling (Chap. XV of [11]) and we act along two directions:

- To explore application areas where nonsmooth optimization algorithms can be applied, possibly after some tayloring. A rich field of such application is combinatorial optimization, with all forms of relaxation [12].
- To explore the possibility of designing more sophisticated algorithms. This implies an appropriate generalization of second derivatives when the first derivative does not exist, and we use advanced tools of nonsmooth analysis, for example [14].

CAGIRE Team

3. Research Program

3.1. Computational fluid mechanics: resolving versus modelling small scales of turbulence

A typical continuous solution of the Navier Stokes equations is governed by a spectrum of time and space scales. The broadness of that spectrum is directly controlled by the Reynolds number defined as the ratio between the inertial forces and the viscous forces. This number is quite helpful to determine if the flow is turbulent or not. In the former case, it indicates the range of scales of fluctuations that are present in the flow under study. Typically, for instance for the velocity field, the ratio between the largest scale (the integral length scale) to the smallest one (Kolmogorov scale) scales as $Re^{3/4}$ per dimension. In addition, for internal flows, the viscous effects near the solid walls yield a scaling proportional to Re per dimension. The smallest scales may have a certain effect on the largest ones which implies that an accurate framework for the modelling and the computation of such turbulent flows must take into account all these scales of time and space fluctuations. This can be achieved either by solving directly the Navier-Stokes (NS) equations (Direct numerical simulations or DNS) or by first applying to them a filtering operation either in time or space. In the latter cases, the closure of the new terms that appear in the filtered equations due to the presence of the non-linear terms implies the recourse to a turbulence model before discretizing and then solving the set of resulting governing equations. Among these different methodologies, the Reynolds averaged Navier-Stokes (RANS) approach yields a system of equations aimed at describing the mean flow properties. The term mean is referring to an ensemble average which is equivalent to a time average only when the flow is statistically stationnary. In that case, the turbulence model aims at expressing the Reynolds stresses either through the solution of dedicated transport equations (second order modelling) or via the recourse to the concept of turbulent viscosity used to write and ad-hoc relation (linear or not) between the Reynolds stress and the mean velocity gradient tensor. If the filtering operation involves a convolution with a filter function in space of width δ, this corresponds to the large-eddy simulation (LES) approach. The structures of size below δ are filtered out while the bigger structures are directly resolved. The resulting set of filtered equations is again not closed and calls for a model aimed at providing a suitable expression for the subgrid scale stress tensor.

From a computational point of view, the RANS approach is the less demanding, which explains why historically it has been the workhorse in both the academic and the industrial sectors. Although it has permitted quite a substantive progress in the understanding of various phenomena such as turbulent combustion or heat transfer, its inability to provide a time-dependent information has led to promote in the last decade the recourse to either LES or DNS as well as hybrid methods that combine RANS and LES. By simulating the large scale structures while modelling the smallest ones supposed to be more isotropic, the LES, alone or combined with the most adanced RANS models such as the EB-RSM model [4] proved to be quite a step through that permits to fully take advantage of the increasing power of computers to study complex flow configurations. In the same time, DNS was progressively applied to geometries of increasing complexity (channel flows, jets, turbulent premixed flames), and proved to be a formidable tool that permits **(i)** to improve our knowledge of turbulent flows and **(ii)** to test (i.e. validate or invalidate) and improve the numerous modelling hypotheses inherently associated to the RANS and LES approaches. From a numerical point of view, if the steady nature of the RANS equations allows to perform iterative convergence on finer and finer meshes, this is no longer possible for LES or DNS which are time-dependent. It is therefore necessary to develop high accuracy schemes in such frameworks. Considering that the Reynolds number in an engine combustion chamber is significantly larger than 10000, a direct numerical simulation of the whole flow domain is not conceivable on a routine basis but the simulation of generic flows which feature some of the phenomena present in a combustion chamber is accessible considering the recent progresses in High Performance Computing (HPC).

3.2. Computational fluid mechanics: numerical methods

All the methods we describe are mesh-based methods: the computational domain is divided into *cells*, that have an elementary shape: triangle and quadrangle in two dimensions, and tetrahedra, hexahedra, pyramids, and prism in three dimensions. If the cells are only regular hexahedra, the mesh is said to be *structured*. Otherwise, it is said to be unstructured. If the mesh is composed of more than one sort of elementary shape, the mesh is said to be *hybrid*.

The basic numerical model for the computation of internal flows is based on the Navier-Stokes equations. For fifty years, many sorts of numerical approximation have been tried for this sort of system: finite differences, finite volumes, and finite elements.

The finite differences have met a great success for some equations, but for the approximation of fluid mechanics, they suffer from two drawbacks. First, structured meshes must be used. This drawback can be very limiting in the context of internal aerodynamics, in which the geometries can be very complex. The other problem is that finite difference schemes do not include any upwinding process, which is essential for convection dominated flows.

The finite volumes methods have imposed themselves in the last thirty years in the context of aerodynamic. They intrinsically contain an upwinding mechanism, so that they are naturally stable for linear as much as for nonlinear convective flows. The extension to diffusive flows has been done in [18]. Whereas the extension to second order with the MUSCL method is widely spread, the extension to higher order has always been a strong drawback of finite volumes methods. For such an extension, reconstruction methods have been developed (ENO, WENO). Nevertheless, these methods need to use a stencil that increases quickly with the order, which induces problems for the parallelisation and the efficiency of the implementation. Another natural extension of finite volume methods are the so-called discontinuous Galerkin methods. These methods are based on the Galerkin' idea of projecting the weak formulation of the equations on a finite dimensional space. But on the contrary to the conforming finite elements method, the approximation space is composed of functions that are continuous (typically: polynomials) inside each cell, but that are discontinuous on the sides. The discontinuous Galerkin methods are currently very popular, because they can be used with many sort of partial differential equations. Moreover, the fact that the approximation is discontinuous allows to use modern mesh adaptation (hanging nodes, which appear in non conforming mesh adaptation), and adaptive order, in which the high order is used only where the solution is smooth.

Discontinuous Galerkin methods where introduced by Reed and Hill [39] and first studied by Lesaint and Raviart [32]. The extension to the Euler system with explicit time integration was mainly led by Shu, Cockburn and their collaborators. The steps of time integration and slope limiting were similar to high order ENO schemes, whereas specific constraints given by the finite elements nature of the scheme were progressively solved, for scalar conservation laws [22], [21], one dimensional systems [20], multidimensional scalar conservation laws [19], and multidimensional systems [23]. For the same system, we can also cite the work of [25], [30], which is slightly different: the stabilisation is made by adding a nonlinear stabilisation term, and the time integration is implicit. Then, the extension to the compressible Navier-Stokes system was made by Bassi and Rebay [17], first by a mixed type finite element method, and then simplified by means of lifting operators. The extension to the $k-\omega$ RANS system was made in [16]. Another type of discontinuous Galerkin method for Navier Stokes is the so-called Symmetric Interior Penalty (SIP) method. It is used for example by Hartmann and Houston [28]. The symmetric nature of the discretization is particularly well suited with mesh adaptation by means of the adjoint equation resolution [29]. Last, we note that the discontinuous Galerkin method was already successfully tested in [24] at Direct Numerical Simulation scale for very moderate Reynolds, and also by the Munz's team in Stuttgart [33], with local time stepping.

For concluding this section, there already exist numerical schemes based on the discontinuous Galerkin method which proved to be efficient for computing compressible viscous flows. Nevertheless, there remain things to be improved, which include: efficient shock capturing term methods for supersonic flows, high order discretization of curved boundaries, or low Mach behaviour of these schemes. Another drawback of the

discontinuous Galerkin methods is that they are very computationally costly, due to the accurate representation of the solution. Accordingly, a particular care must be taken on the implementation for being efficient.

3.3. Flow analysis and CFD assessment: experimental aspects

The capability of producing in-situ experimental data is another originality of our project. By carefully controlling the flow configuration and the type of data we are measuring, we are in situation of assessing in depth the quality of our simulations results over the complete spectrum of possible approaches ranging from DNS, RANS and Hybrid RANS-LES models that the team is developing or LES.

The flow configuration we have chosen is that of a jet in cross-flow since it features large scale coherent structures, flow separation, turbulence and wall-flow interaction.

A great deal of experiments has been devoted to the study of jet in crossflow configurations. They essentially differ one from each other by the hole shape (cylindrical or shaped), the hole axis inclination, the way by which the hole is fed, the characteristics of the crossflow and the jet (turbulent or not, isothermal or not), the number of holes considered and last but not least the techniques used to investigate the flow. A good starting point to assess the diversity of the studies carried out is given by [34]. For inclined cylindrical holes, the experimental database produced by Gustafsson and Johansson [0] represents a sound reference base and for normal injection, the work by [40] served as reference for LES simulations [38]. For shaped holes, the studies are less numerous and are aimed at assessing the influence of the hole shape on various flow properties such as the heat transfer at the wall [31]. In 2007, Most [35] developed at UPPA a test facility for studying jet in crossflow issued from shaped holes. The hole shape was chosen as a 12.5 scale of the holes (i.e. at scale 1) drilled by laser in a combustion chamber. His preliminary 2-component PIV results have been used to test RANS simulations [36] and LES [37]. Later, in the framework of the KIAI FP7 European programme, Florenciano [26] upgraded the rig by implementing an acoustic forcing device of the crossflow stream and by performing phase-locked PIV measurements that were used to test the accuracy of LES results. Thus, this test facility is extensively used in the framework of the present project to investigate a 1-hole cylindrical inclined jet interacting with a turbulent crossflow. PIV and LDV metrology are the workhorses as far as metrology is concerned.

[0] Slanted jet

CLASSIC Project-Team

3. Research Program

3.1. Regression models of supervised learning

The most obvious contribution of statistics to machine learning is to consider the supervised learning scenario as a special case of regression estimation: given n independent pairs of observations (X_i, Y_i), $i = 1, \cdots, n$, the aim is to "learn" the dependence of Y_i on X_i. Thus, classical results about statistical regression estimation apply, with the caveat that the hypotheses we can reasonably assume about the distribution of the pairs (X_i, Y_i) are much weaker than what is usually considered in statistical studies. The aim here is to assume very little, maybe only independence of the observed sequence of input-output pairs, and to validate model and variable selection schemes. These schemes should produce the best possible approximation of the joint distribution of (X_i, Y_i) within some restricted family of models. Their performance is evaluated according to some measure of discrepancy between distributions, a standard choice being to use the Kullback-Leibler divergence.

3.1.1. PAC-Bayes inequalities

One of the specialties of the team in this direction is to use PAC-Bayes inequalities to combine thresholded exponential moment inequalities. The name of this theory comes from its founder, David McAllester, and may be misleading. Indeed, its cornerstone is rather made of non-asymptotic entropy inequalities, and a perturbative approach to parameter estimation. The team has made major contributions to the theory, first focussed on classification [6], then on regression [1] and on principal component analysis of a random sample of points in high dimension. It has introduced the idea of combining the PAC-Bayesian approach with the use of thresholded exponential moments [7], in order to derive bounds under very weak assumptions on the noise.

COMMANDS Project-Team

3. Research Program

3.1. Historical aspects

The roots of deterministic optimal control are the "classical" theory of the calculus of variations, illustrated by the work of Newton, Bernoulli, Euler, and Lagrange (whose famous multipliers were introduced in [84]), with improvements due to the "Chicago school", Bliss [51] during the first part of the 20th century, and by the notion of relaxed problem and generalized solution (Young [93]).

Trajectory optimization really started with the spectacular achievement done by Pontryagin's group [90] during the fifties, by stating, for general optimal control problems, nonlocal optimality conditions generalizing those of Weierstrass. This motivated the application to many industrial problems (see the classical books by Bryson and Ho [59], Leitmann [86], Lee and Markus [85], Ioffe and Tihomirov [76]). Since then, various theoretical achievements have been obtained by extending the results to nonsmooth problems, see Aubin [47], Clarke [60], Ekeland [67].

Dynamic programming was introduced and systematically studied by R. Bellman during the fifties. The HJB equation, whose solution is the value function of the (parameterized) optimal control problem, is a variant of the classical Hamilton-Jacobi equation of mechanics for the case of dynamics parameterized by a control variable. It may be viewed as a differential form of the dynamic programming principle. This nonlinear first-order PDE appears to be well-posed in the framework of *viscosity solutions* introduced by Crandall and Lions [62], [63], [61]. These tools also allow to perform the numerical analysis of discretization schemes. The theoretical contributions in this direction did not cease growing, see the books by Barles [49] and Bardi and Capuzzo-Dolcetta [48].

3.2. Trajectory optimization

The so-called *direct methods* consist in an optimization of the trajectory, after having discretized time, by a nonlinear programming solver that possibly takes into account the dynamic structure. So the two main problems are the choice of the discretization and the nonlinear programming algorithm. A third problem is the possibility of refinement of the discretization once after solving on a coarser grid.

In the *full discretization approach*, general Runge-Kutta schemes with different values of control for each inner step are used. This allows to obtain and control high orders of precision, see Hager [73], Bonnans [54]. In an interior-point algorithm context, controls can be eliminated and the resulting system of equation is easily solved due to its band structure. Discretization errors due to constraints are discussed in Dontchev et al. [66]. See also Malanowski et al. [87].

In the *indirect* approach, the control is eliminated thanks to Pontryagin's maximum principle. One has then to solve the two-points boundary value problem (with differential variables state and costate) by a single or multiple shooting method. The questions are here the choice of a discretization scheme for the integration of the boundary value problem, of a (possibly globalized) Newton type algorithm for solving the resulting finite dimensional problem in $I\!R^n$ (n is the number of state variables), and a methodology for finding an initial point.

For state constrained problems or singular arcs, the formulation of the shooting function may be quite elaborate [52], [53], [46]. As initiated in [70], we focus more specifically on the handling of discontinuities, with ongoing work on the geometric integration aspects (Hamiltonian conservation).

3.3. Hamilton-Jacobi-Bellman approach

This approach consists in calculating the value function associated with the optimal control problem, and then synthesizing the feedback control and the optimal trajectory using Pontryagin's principle. The method has the great particular advantage of reaching directly the global optimum, which can be very interesting when the problem is not convex.

Characterization of the value function >From the dynamic programming principle, we derive a characterization of the value function as being a solution (in viscosity sense) of an Hamilton-Jacobi-Bellman equation, which is a nonlinear PDE of dimension equal to the number n of state variables. Since the pioneer works of Crandall and Lions [62], [63], [61], many theoretical contributions were carried out, allowing an understanding of the properties of the value function as well as of the set of admissible trajectories. However, there remains an important effort to provide for the development of effective and adapted numerical tools, mainly because of numerical complexity (complexity is exponential with respect to n).

Numerical approximation for continuous value function Several numerical schemes have been already studied to treat the case when the solution of the HJB equation (the value function) is continuous. Let us quote for example the Semi-Lagrangian methods [69], [68] studied by the team of M. Falcone (La Sapienza, Rome), the high order schemes WENO, ENO, Discrete galerkin introduced by S. Osher, C.-W. Shu, E. Harten [74], [75], [75], [88], and also the schemes on nonregular grids by R. Abgrall [45], [44]. All these schemes rely on finite differences or/and interpolation techniques which lead to numerical diffusions. Hence, the numerical solution is unsatisfying for long time approximations even in the continuous case.

One of the (nonmonotone) schemes for solving the HJB equation is based on the Ultrabee algorithm proposed, in the case of advection equation with constant velocity, by Roe [92] and recently revisited by Després-Lagoutière [65], [64]. The numerical results on several academic problems show the relevance of the antidiffusive schemes. However, the theoretical study of the convergence is a difficult question and is only partially done.

Optimal stochastic control problems occur when the dynamical system is uncertain. A decision typically has to be taken at each time, while realizations of future events are unknown (but some information is given on their distribution of probabilities). In particular, problems of economic nature deal with large uncertainties (on prices, production and demand). Specific examples are the portfolio selection problems in a market with risky and non-risky assets, super-replication with uncertain volatility, management of power resources (dams, gas). Air traffic control is another example of such problems.

Nonsmoothness of the value function. Sometimes the value function is smooth (e.g. in the case of Merton's portfolio problem, Oksendal [94]) and the associated HJB equation can be solved explicitly. Still, the value function is not smooth enough to satisfy the HJB equation in the classical sense. As for the deterministic case, the notion of viscosity solution provides a convenient framework for dealing with the lack of smoothness, see Pham [89], that happens also to be well adapted to the study of discretization errors for numerical discretization schemes [77], [50].

Numerical approximation for optimal stochastic control problems. The numerical discretization of second order HJB equations was the subject of several contributions. The book of Kushner-Dupuis [83] gives a complete synthesis on the Markov chain schemes (i.e Finite Differences, semi-Lagrangian, Finite Elements, ...). Here a main difficulty of these equations comes from the fact that the second order operator (i.e. the diffusion term) is not uniformly elliptic and can be degenerated. Moreover, the diffusion term (covariance matrix) may change direction at any space point and at any time (this matrix is associated the dynamics volatility).

For solving stochastic control problems, we studied the so-called Generalized Finite Differences (GFD), that allow to choose at any node, the stencil approximating the diffusion matrix up to a certain threshold [57]. Determining the stencil and the associated coefficients boils down to a quadratic program to be solved at each point of the grid, and for each control. This is definitely expensive, with the exception of special structures where the coefficients can be computed at low cost. For two dimensional systems, we designed a (very) fast algorithm for computing the coefficients of the GFD scheme, based on the Stern-Brocot tree [56].

CORIDA Team

3. Research Program

3.1. Analysis and control of fluids and of fluid-structure interactions

Participants: Thomas Chambrion, Antoine Henrot, Alexandre Munnier, Lionel Rosier, Jean-François Scheid, Takéo Takahashi, Marius Tucsnak, Jean-Claude Vivalda.

The problems we consider are modeled by the Navier-Stokes, Euler or Korteweg de Vries equations (for the fluid) coupled to the equations governing the motion of the solids. One of the main difficulties of this problem comes from the fact that the domain occupied by the fluid is one of the unknowns of the problem. We have thus to tackle a *free boundary problem*.

The control of fluid flows is a major challenge in many applications: aeronautics, pollution issues, regulation of irrigation channels or of the flow in pipelines, etc. All these problems cannot be easily reduced to finite dimensional models so a methodology of analysis and control based on PDE's is an essential issue. In a first approximation the motion of fluid and of the solids can be decoupled. The most used models for an incompressible fluid are given by the Navier-Stokes or by the Euler equations.

The optimal open loop control approach of these models has been developed from both the theoretical and numerical points of view. Controllability issues for the equations modeling the fluid motion are by now well understood (see, for instance, Imanuvilov [52] and the references therein). The feedback control of fluid motion has also been recently investigated by several research teams (see, for instance Barbu [47] and references therein) but this field still contains an important number of open problems (in particular those concerning observers and implementation issues). One of our aims is to develop efficient tools for computing feedback laws for the control of fluid systems.

In real applications the fluid is often surrounded by or it surrounds an elastic structure. In the above situation one has to study fluid-structure interactions. This subject has been intensively studied during the last years, in particular for its applications in noise reduction problems, in lubrication issues or in aeronautics. In this kind of problems, a PDE's system modeling the fluid in a cavity (Laplace equation, wave equation, Stokes, Navier-Stokes or Euler systems) is coupled to the equations modeling the motion of a part of the boundary. The difficulties of this problem are due to several reasons such as the strong nonlinear coupling and the existence of a free boundary. This partially explains the fact that applied mathematicians have only recently tackled these problems from either the numerical or theoretical point of view. One of the main results obtained in our project concerns the global existence of weak solutions in the case of a two-dimensional Navier–Stokes fluid [59]. Another important result gives the existence and the uniqueness of strong solutions for two or three-dimensional Navier–Stokes fluid [61]. In that case, the solution exists as long as there is no contact between rigid bodies, and for small data in the three-dimensional case.

3.2. Frequency domain methods for the analysis and control of systems governed by PDE's

Participants: Xavier Antoine, Bruno Pinçon, Karim Ramdani.

We use frequency tools to analyze different types of problems. The first one concerns the control, the optimal control and the stabilization of systems governed by PDE's, and their numerical approximations. The second one concerns time-reversal phenomena, while the last one deals with numerical approximation of high-frequency scattering problems.

3.2.1. Control and stabilization for skew-adjoint systems

The first area concerns theoretical and numerical aspects in the control of a class of PDE's. More precisely, in a semigroup setting, the systems we consider have a skew-adjoint generator. Classical examples are the wave, the Bernoulli-Euler or the Schrödinger equations. Our approach is based on an original characterization of exact controllability of second order conservative systems proposed by K. Liu [56]. This characterization can be related to the Hautus criterion in the theory of finite dimensional systems (cf. [51]). It provides for time-dependent problems exact controllability criteria *that do not depend on time, but depend on the frequency variable* conjugated to time. Studying the controllability of a given system amounts then to establishing uniform (with respect to frequency) estimates. In other words, the problem of exact controllability for the wave equation, for instance, comes down to a high-frequency analysis for the Helmholtz operator. This frequency approach has been proposed first by K. Liu for bounded control operators (corresponding to internal control problems), and has been recently extended to the case of unbounded control operators (and thus including boundary control problems) by L. Miller [57]. Using the result of Miller, K. Ramdani, T. Takahashi, M. Tucsnak have obtained in [5] a new spectral formulation of the criterion of Liu [56], which is valid for boundary control problems. This frequency test can be seen as an observability condition for packets of eigenvectors of the operator. This frequency test has been successfully applied in [5] to study the exact controllability of the Schrödinger equation, the plate equation and the wave equation in a square. Let us emphasize here that one further important advantage of this frequency approach lies in the fact that it can also be used for the analysis of space semi-discretized control problems (by finite element or finite differences). The estimates to be proved must then be uniform with respect to *both the frequency and the mesh size*.

In the case of finite dimensional systems one of the main applications of frequency domain methods consists in designing robust controllers, in particular of H^∞ type. Obtaining the similar tools for systems governed by PDE's is one of the major challenges in the theory of infinite dimensional systems. The first difficulty which has to be tackled is that, even for very simple PDE systems, no method giving the parametrisation of all stabilizing controllers is available. One of the possible remedies consists in considering known families of stabilizing feedback laws depending on several parameters and in optimizing the H^∞ norm of an appropriate transfer function with respect to this parameters. Such families of feedback laws yielding computationally tractable optimization problems are now available for systems governed by PDE's in one space dimension.

3.2.2. Time-reversal

The second area in which we make use of frequency tools is the analysis of time-reversal for harmonic acoustic waves. This phenomenon described in Fink [49] is a direct consequence of the reversibility of the wave equation in a non dissipative medium. It can be used to **focus an acoustic wave** on a target through a complex and/or unknown medium. To achieve this, the procedure followed is quite simple. First, time-reversal mirrors are used to generate an incident wave that propagates through the medium. Then, the mirrors measure the acoustic field diffracted by the targets, time-reverse it and back-propagate it in the medium. Iterating the scheme, we observe that the incident wave emitted by the mirrors focuses on the scatterers. An alternative and more original focusing technique is based on the so-called D.O.R.T. method [50]. According to this experimental method, the eigenelements of the time-reversal operator contain important information on the propagation medium and on the scatterers contained in it. More precisely, the number of nonzero eigenvalues is exactly the number of scatterers, while each eigenvector corresponds to an incident wave that selectively focuses on each scatterer.

Time-reversal has many applications covering a wide range of fields, among which we can cite *medicine* (kidney stones destruction or medical imaging), *sub-marine communication* and *non destructive testing*. Let us emphasize that in the case of time-harmonic acoustic waves, time-reversal is equivalent to phase conjugation and involves the Helmholtz operator.

In [2], we proposed the first far field model of time reversal in the time-harmonic case.

3.2.3. Numerical approximation of high-frequency scattering problems

This subject deals mainly with the numerical solution of the Helmholtz or Maxwell equations for open region scattering problems. This kind of situation can be met e.g. in radar systems in electromagnetism or in acoustics for the detection of underwater objects like submarines.

Two particular difficulties are considered in this situation

- the wavelength of the incident signal is small compared to the characteristic size of the scatterer,
- the problem is set in an unbounded domain.

These two problematics limit the application range of most common numerical techniques. The aim of this part is to develop new numerical simulation techniques based on microlocal analysis for modeling the propagation of rays. The importance of microlocal techniques in this situation is that it makes possible a local analysis both in the spatial and frequency domain. Therefore, it can be seen as a kind of asymptotic theory of rays which can be combined with numerical approximation techniques like boundary element methods. The resulting method is called the On-Surface Radiation Condition method.

3.3. Observability, controllability and stabilization in the time domain

Participants: Fatiha Alabau-Boussouira, Xavier Antoine, Thomas Chambrion, Antoine Henrot, Karim Ramdani, Marius Tucsnak, Jean-Claude Vivalda.

Controllability and observability have been set at the center of control theory by the work of R. Kalman in the 1960's and soon they have been generalized to the infinite-dimensional context. The main early contributors have been D.L. Russell, H. Fattorini, T. Seidman, R. Triggiani, W. Littman and J.-L. Lions. The latter gave the field an enormous impact with his book [54], which is still a main source of inspiration for many researchers. Unlike in classical control theory, for infinite-dimensional systems there are many different (and not equivalent) concepts of controllability and observability. The strongest concepts are called exact controllability and exact observability, respectively. In the case of linear systems exact controllability is important because it guarantees stabilizability and the existence of a linear quadratic optimal control. Dually, exact observability guarantees the existence of an exponentially converging state estimator and the existence of a linear quadratic optimal filter. An important feature of infinite dimensional systems is that, unlike in the finite dimensional case, the conditions for exact observability are no longer independent of time. More precisely, for simple systems like a string equation, we have exact observability only for times which are large enough. For systems governed by other PDE's (like dispersive equations) the exact observability in arbitrarily small time has been only recently established by using new frequency domain techniques. A natural question is to estimate the energy required to drive a system in the desired final state when the control time goes to zero. This is a challenging theoretical issue which is critical for perturbation and approximation problems. In the finite dimensional case this issue has been first investigated in Seidman [60]. In the case of systems governed by linear PDE's some similar estimates have been obtained only very recently (see, for instance Miller [57]). One of the open problems of this field is to give sharp estimates of the observability constants when the control time goes to zero.

Even in the finite-dimensional case, despite the fact that the linear theory is well established, many challenging questions are still open, concerning in particular nonlinear control systems.

In some cases it is appropriate to regard external perturbations as unknown inputs; for these systems the synthesis of observers is a challenging issue, since one cannot take into account the term containing the unknown input into the equations of the observer. While the theory of observability for linear systems with unknown inputs is well established, this is far from being the case in the nonlinear case. A related active field of research is the uniform stabilization of systems with time-varying parameters. The goal in this case is to stabilize a control system with a control strategy independent of some signals appearing in the dynamics, i.e., to stabilize simultaneously a family of time-dependent control systems and to characterize families of control systems that can be simultaneously stabilized.

One of the basic questions in finite- and infinite-dimensional control theory is that of motion planning, i.e., the explicit design of a control law capable of driving a system from an initial state to a prescribed final one. Several techniques, whose suitability depends strongly on the application which is considered, have been and are being developed to tackle such a problem, as for instance the continuation method, flatness, tracking or optimal control. Preliminary to any question regarding motion planning or optimal control is the issue of controllability, which is not, in the general nonlinear case, solved by the verification of a simple algebraic criterion. A further motivation to study nonlinear controllability criteria is given by the fact that techniques developed in the domain of (finite-dimensional) geometric control theory have been recently applied successfully to study the controllability of infinite-dimensional control systems, namely the Navier–Stokes equations (see Agrachev and Sarychev [46]).

3.4. Implementation

This is a transverse research axis since all the research directions presented above have to be validated by giving control algorithms which are aimed to be implemented in real control systems. We stress below some of the main points which are common (from the implementation point of view) to the application of the different methods described in the previous sections.

For many infinite dimensional systems the use of co-located actuators and sensors and of simple proportional feed-back laws gives satisfying results. However, for a large class of systems of interest it is not clear that these feedbacks are efficient, or the use of co-located actuators and sensors is not possible. This is why a more general approach for the design of the feedbacks has to be considered. Among the techniques in finite dimensional systems theory those based on the solutions of infinite dimensional Riccati equation seem the most appropriate for a generalization to infinite dimensional systems. The classical approach is to approximate an LQR problem for a given infinite dimensional system by finite dimensional LQR problems. As it has been already pointed out in the literature this approach should be carefully analyzed since, even for some very simple examples, the sequence of feedbacks operators solving the finite dimensional LQR is not convergent. Roughly speaking this means that by refining the mesh we obtain a closed loop system which is not exponentially stable (even if the corresponding infinite dimensional system is theoretically stabilized). In order to overcome this difficulty, several methods have been proposed in the literature : filtering of high frequencies, multigrid methods or the introduction of a numerical viscosity term. We intend to first apply the numerical viscosity method introduced in Tcheougoue Tebou – Zuazua [62], for optimal and robust control problems.

CQFD Project-Team

3. Research Program

3.1. Introduction

The scientific objectives of the team are to provide mathematical tools for modeling and optimization of complex systems. These systems require mathematical representations which are in essence dynamic, multi-model and stochastic. This increasing complexity poses genuine scientific challenges in the domain of modeling and optimization. More precisely, our research activities are focused on stochastic optimization and (parametric, semi-parametric, multidimensional) statistics which are complementary and interlinked topics. It is essential to develop simultaneously statistical methods for the estimation and control methods for the optimization of the models.

3.2. Main research topics

- Stochastic modeling: Markov chain, Piecewise Deterministic Markov Processes (PDMP), Markov Decision Processes (MDP).
  The mathematical representation of complex systems is a preliminary step to our final goal corresponding to the optimization of its performance. For example, in order to optimize the predictive maintenance of a system, it is necessary to choose the adequate model for its representation. The step of modeling is crucial before any estimation or computation of quantities related to its optimization. For this we have to represent all the different regimes of the system and the behavior of the physical variables under each of these regimes. Moreover, we must also select the dynamic variables which have a potential effect on the physical variable and the quantities of interest. The team CQFD works on the theory of Piecewise Deterministic Markov Processes (PDMP's) and on Markov Decision Processes (MDP's). These two classes of systems form general families of controlled stochastic processes suitable for the modeling of sequential decision-making problems in the continuous-time (PDMPs) and discrete-time (MDP's) context. They appear in many fields such as engineering, computer science, economics, operations research and constitute powerful class of processes for the modeling of complex system.
- Estimation methods: estimation for PDMP; estimation in non- and semi parametric regression modeling.
  To the best of our knowledge, there does not exist any general theory for the problems of estimating parameters of PDMPs although there already exist a large number of tools for sub-classes of PDMPs such as point processes and marked point processes. However, to fill the gap between these specific models and the general class of PDMPs, new theoretical and mathematical developments will be on the agenda of the whole team. In the framework of non-parametric regression or quantile regression, we focus on kernel estimators or kernel local linear estimators for complete data or censored data. New strategies for estimating semi-parametric models via recursive estimation procedures have also received an increasing interest recently. The advantage of the recursive estimation approach is to take into account the successive arrivals of the information and to refine, step after step, the implemented estimation algorithms. These recursive methods do require restarting calculation of parameter estimation from scratch when new data are added to the base. The idea is to use only the previous estimations and the new data to refresh the estimation. The gain in time could be very interesting and there are many applications of such approaches.

- Dimension reduction: dimension-reduction via SIR and related methods, dimension-reduction via multidimensional and classification methods.
  Most of the dimension reduction approaches seek for lower dimensional subspaces minimizing the loss of some statistical information. This can be achieved in modeling framework or in exploratory data analysis context.
  In modeling framework we focus our attention on semi-parametric models in order to conjugate the advantages of parametric and nonparametric modeling. On the one hand, the parametric part of the model allows a suitable interpretation for the user. On the other hand, the functional part of the model offers a lot of flexibility. In this project, we are especially interested in the semi-parametric regression model $Y = f(X'\theta) + \varepsilon$, the unknown parameter θ belongs to $\mathbb{R}^p$ for a single index model, or is such that $\theta = [\theta_1, \cdots, \theta_d]$ (where each θ_k belongs to $\mathbb{R}^p$ and $d \leq p$ for a multiple indices model), the noise ε is a random error with unknown distribution, and the link function f is an unknown real valued function. Another way to see this model is the following: the variables X and Y are independent given $X'\theta$. In our semi-parametric framework, the main objectives are to estimate the parametric part θ as well as the nonparametric part which can be the link function f, the conditional distribution function of Y given X or the conditional quantile q_α. In order to estimate the dimension reduction parameter θ we focus on the Sliced Inverse Regression (SIR) method which has been introduced by Li [57] and Duan and Li [55]
  Methods of dimension reduction are also important tools in the field of data analysis, data mining and machine learning.They provide a way to understand and visualize the structure of complex data sets.Traditional methods among others are principal component analysis for quantitative variables or multiple component analysis for qualitative variables. New techniques have also been proposed to address these challenging tasks involving many irrelevant and redundant variables and often comparably few observation units. In this context, we focus on the problem of synthetic variables construction, whose goals include increasing the predictor performance and building more compact variables subsets. Clustering of variables is used for feature construction. The idea is to replace a group of "similar" variables by a cluster centroid, which becomes a feature. The most popular algorithms include K-means and hierarchical clustering. For a review, see, e.g., the textbook of Duda [56]
- Stochastic optimal control: optimal stopping, impulse control, continuous control, linear programming.
  The first objective is to focus on the development of computational methods.
  - In the continuous-time context, stochastic control theory has from the numerical point of view, been mainly concerned with Stochastic Differential Equations (SDEs in short). From the practical and theoretical point of view, the numerical developments for this class of processes are extensive and largely complete. It capitalizes on the connection between SDEs and second order partial differential equations (PDEs in short) and the fact that the properties of the latter equations are very well understood. It is, however, hard to deny that the development of computational methods for the control of PDMPs has received little attention. One of the main reasons is that the role played by the familiar PDEs in the diffusion models is here played by certain systems of integro-differential equations for which there is not (and cannot be) a unified theory such as for PDEs as emphasized by M.H.A. Davis in his book. To the best knowledge of the team, there is only one attempt to tackle this difficult problem by O.L.V. Costa and M.H.A. Davis. The originality of our project consists in studying this unexplored area. It is very important to stress the fact that these numerical developments will give rise to a lot of theoretical issues such as type of approximations, convergence results, rates of convergence,....
  - Theory for MDP's has reached a rather high degree of maturity, although the classical tools such as value iteration, policy iteration and linear programming, and their various extensions, are not applicable in practice. We believe that the theoretical progress of MDP's must be in parallel with the corresponding numerical developments. Therefore, solving

MDP's numerically is an awkward and important problem both from the theoretical and practical point of view. In order to meet this challenge, the fields of neural networks, neuro-dynamic programming and approximate dynamic programming became recently an active area of research. Such methods found their roots in heuristic approaches, but theoretical results for convergence results are mainly obtained in the context of finite MDP's. Hence, an ambitious challenge is to investigate such numerical problems but for models with general state and action spaces. Our motivation is to develop theoretically consistent computational approaches for approximating optimal value functions and finding optimal policies.

– An effort has been devoted to the development of efficient computational methods in the setting of communication networks. These are complex dynamical systems composed of several interacting nodes that exhibit important congestion phenomena as their level of interaction grows. The dynamics of such systems are affected by the randomness of their underlying events (e.g., arrivals of http requests to a web-server) and are described stochastically in terms of queueing network models. These are mathematical tools that allow one to predict the performance achievable by the system, to optimize the network configuration, to perform capacity-planning studies, etc. These objectives are usually difficult to achieve without a mathematical model because Internet systems are huge in size. However, because of the exponential growth of their state spaces, an exact analysis of queueing network models is generally difficult to obtain. Given this complexity, we have developed analyses in some limiting regime of practical interest (e.g., systems size grows to infinity). This approach is helpful to obtain a simpler mathematical description of the system under investigation, which leads to the direct definition of efficient, though approximate, computational methods and also allows to investigate other aspects such as Nash equilibria.

The second objective of the team is to study some theoretical aspects related to MDPs such as convex analytical methods and singular perturbation. Analysis of various problems arising in MDPs leads to a large variety of interesting mathematical problems.

DEFI Project-Team

3. Research Program

3.1. Research Program

The research activity of our team is dedicated to the design, analysis and implementation of efficient numerical methods to solve inverse and shape/topological optimization problems in connection with wave imaging, structural design, non-destructive testing and medical imaging modalities. We are particularly interested in the development of fast methods that are suited for real-time applications and/or large scale problems. These goals require to work on both the physical and the mathematical models involved and indeed a solid expertise in related numerical algorithms.

This section intends to give a general overview of our research interests and themes. We choose to present them through the specific academic example of inverse scattering problems (from inhomogeneities), which is representative of foreseen developments on both inversion and (topological) optimization methods. The practical problem would be to identify an inclusion from measurements of diffracted waves that result from the interaction of the sought inclusion with some (incident) waves sent into the probed medium. Typical applications include biomedical imaging where using micro-waves one would like to probe the presence of pathological cells, or imaging of urban infrastructures where using ground penetrating radars (GPR) one is interested in finding the location of buried facilities such as pipelines or waste deposits. This kind of applications requires in particular fast and reliable algorithms.

By “imaging” we shall refer to the inverse problem where the concern is only the location and the shape of the inclusion, while “identification” may also indicate getting informations on the inclusion physical parameters.

Both problems (imaging and identification) are non linear and ill-posed (lack of stability with respect to measurements errors if some careful constrains are not added). Moreover, the unique determination of the geometry or the coefficients is not guaranteed in general if sufficient measurements are not available. As an example, in the case of anisotropic inclusions, one can show that an appropriate set of data uniquely determine the geometry but not the material properties.

These theoretical considerations (uniqueness, stability) are not only important in understanding the mathematical properties of the inverse problem, but also guide the choice of appropriate numerical strategies (which information can be stably reconstructed) and also the design of appropriate regularization techniques. Moreover, uniqueness proofs are in general constructive proofs, i.e. they implicitly contain a numerical algorithm to solve the inverse problem, hence their importance for practical applications. The sampling methods introduced below are one example of such algorithms.

A large part of our research activity is dedicated to numerical methods applied to the first type of inverse problems, where only the geometrical information is sought. In its general setting the inverse problem is very challenging and no method can provide a universal satisfactory solution to it (regarding the balance cost-precision-stability). This is why in the majority of the practically employed algorithms, some simplification of the underlying mathematical model is used, according to the specific configuration of the imaging experiment. The most popular ones are geometric optics (the Kirchhoff approximation) for high frequencies and weak scattering (the Born approximation) for small contrasts or small obstacles. They actually give full satisfaction for a wide range of applications as attested by the large success of existing imaging devices (radar, sonar, ultrasound, X-ray tomography, etc.), that rely on one of these approximations.

Generally speaking, the used simplifications result in a linearization of the inverse problem and therefore are usually valid only if the latter is weakly non-linear. The development of these simplified models and the improvement of their efficiency is still a very active research area. With that perspective we are particularly interested in deriving and studying higher order asymptotic models associated with small geometrical parameters such as: small obstacles, thin coatings, wires, periodic media, ... Higher order models usually introduce some non linearity in the inverse problem, but are in principle easier to handle from the numerical point of view than in the case of the exact model.

A larger part of our research activity is dedicated to algorithms that avoid the use of such approximations and that are efficient where classical approaches fail: i.e. roughly speaking when the non linearity of the inverse problem is sufficiently strong. This type of configuration is motivated by the applications mentioned below, and occurs as soon as the geometry of the unknown media generates non negligible multiple scattering effects (multiply-connected and closely spaces obstacles) or when the used frequency is in the so-called resonant region (wave-length comparable to the size of the sought medium). It is therefore much more difficult to deal with and requires new approaches. Our ideas to tackle this problem will be motivated and inspired by recent advances in shape and topological optimization methods and also the introduction of novel classes of imaging algorithms, so-called sampling methods.

The sampling methods are fast imaging solvers adapted to multi-static data (multiple receiver-transmitter pairs) at a fixed frequency. Even if they do not use any linearization the forward model, they rely on computing the solutions to a set of linear problems of small size, that can be performed in a completely parallel procedure. Our team has already a solid expertise in these methods applied to electromagnetic 3-D problems. The success of such approaches was their ability to provide a relatively quick algorithm for solving 3-D problems without any need for a priori knowledge on the physical parameters of the targets. These algorithms solve only the imaging problem, in the sense that only the geometrical information is provided.

Despite the large efforts already spent in the development of this type of methods, either from the algorithmic point of view or the theoretical one, numerous questions are still open. These attractive new algorithms also suffer from the lack of experimental validations, due to their relatively recent introduction. We also would like to invest on this side by developing collaborations with engineering research groups that have experimental facilities. From the practical point of view, the most potential limitation of sampling methods would be the need of a large amount of data to achieve a reasonable accuracy. On the other hand, optimization methods do not suffer from this constrain but they require good initial guess to ensure convergence and reduce the number of iterations. Therefore it seems natural to try to combine the two class of methods in order to calibrate the balance between cost and precision.

Among various shape optimization methods, the Level Set method seems to be particularly suited for such a coupling. First, because it shares similar mechanism as sampling methods: the geometry is captured as a level set of an "indicator function" computed on a cartesian grid. Second, because the two methods do not require any a priori knowledge on the topology of the sought geometry. Beyond the choice of a particular method, the main question would be to define in which way the coupling can be achieved. Obvious strategies consist in using one method to pre-process (initialization) or post-process (find the level set) the other. But one can also think of more elaborate ones, where for instance a sampling method can be used to optimize the choice of the incident wave at each iteration step.The latter point is closely related to the design of so called "focusing incident waves" (which are for instance the basis of applications of the time-reversal principle). In the frequency regime, these incident waves can be constructed from the eigenvalue decomposition of the data operator used by sampling methods. The theoretical and numerical investigations of these aspects are still not completely understood for electromagnetic or elastodynamic problems.

Other topological optimization methods, like the homogenization method or the topological gradient method, can also be used, each one provides particular advantages in specific configurations. It is evident that the development of these methods is very suited to inverse problems and provide substantial advantage compared to classical shape optimization methods based on boundary variation. Their applications to inverse problems has not been fully investigated. The efficiency of these optimization methods can also be increased for adequate asymptotic configurations. For instance small amplitude homogenization method can be used as an efficient relaxation method for the inverse problem in the presence of small contrasts. On the other hand, the topological gradient method has shown to perform well in localizing small inclusions with only one iteration.

A broader perspective would be the extension of the above mentioned techniques to time-dependent cases. Taking into account data in time domain is important for many practical applications, such as imaging in cluttered media, the design of absorbing coatings or also crash worthiness in the case of structural design.

For the identification problem, one would like to also have information on the physical properties of the targets. Of course optimization methods is a tool of choice for these problems. However, in some applications

only a qualitative information is needed and obtaining it in a cheaper way can be performed using asymptotic theories combined with sampling methods. We also refer here to the use of so called transmission eigenvalues as qualitative indicators for non destructive testing of dielectrics.

We are also interested in parameter identification problems arising in diffusion-type problems. Our research here is mostly motivated by applications to the imaging of biological tissues with the technique of Diffusion Magnetic Resonance Imaging (DMRI). Roughly speaking DMRI gives a measure of the average distance travelled by water molecules in a certain medium and can give useful information on cellular structure and structural change when the medium is biological tissue. In particular, we would like to infer from DMRI measurements changes in the cellular volume fraction occurring upon various physiological or pathological conditions as well as the average cell size in the case of tumor imaging. The main challenges here are 1) correctly model measured signals using diffusive-type time-dependent PDEs 2) numerically handle the complexity of the tissues 3) use the first two to identify physically relevant parameters from measurements. For the last point we are particularly interested in constructing reduced models of the multiple-compartment Bloch-Torrey partial differential equation using homogenization methods.

DISCO Project-Team

3. Research Program

3.1. Modeling of complex environment

We want to model phenomena such as a temporary loss of connection (e.g. synchronisation of the movements through haptic interfaces), a nonhomogeneous environment (e.g. case of cryogenic systems) or the presence of the human factor in the control loop (e.g. grid systems) but also problems involved with technological constraints (e.g. range of the sensors). The mathematical models concerned include integro-differential, partial differential equations, algebraic inequalities with the presence of several time scales, whose variables and/or parameters must satisfy certain constraints (for instance, positivity).

3.2. Analysis of interconnected systems

- Algebraic analysis of linear systems

  Study of the structural properties of linear differential time-delay systems and linear infinite-dimensional systems (e.g. invariants, controllability, observability, flatness, reductions, decomposition, decoupling, equivalences) by means of constructive algebra, module theory, homological algebra, algebraic analysis and symbolic computation [8], [9], [71], [91], [72], [75].

- Robust stability of linear systems

  Within an interconnection context, lots of phenomena are modelled directly or after an approximation by delay systems. These systems might have fixed delays, time-varying delays, distributed delays ...

  For various infinite-dimensional systems, particularly delay and fractional systems, input-output and time-domain methods are jointly developed in the team to characterize stability. This research is developed at four levels: analytic approaches (H_∞-stability, BIBO-stablity, robust stability, robustness metrics) [1], [2], [5], [6], symbolic computation approaches (SOS methods are used for determining easy-to-check conditions which guarantee that the poles of a given linear system are not in the closed right half-plane, certified CAD techniques), numerical approaches (root-loci, continuation methods) and by means of softwares developed in the team [5], [6].

- Robustness/fragility of biological systems

  Deterministic biological models describing, for instance, species interactions, are frequently composed of equations with important disturbances and poorly known parameters. To evaluate the impact of the uncertainties, we use the techniques of designing of global strict Lyapunov functions or functional developed in the team.

  However, for other biological systems, the notion of robustness may be different and this question is still in its infancy (see, e.g. [83]). Unlike engineering problems where a major issue is to maintain stability in the presence of disturbances, a main issue here is to maintain the system response in the presence of disturbances. For instance, a biological network is required to keep its functioning in case of a failure of one of the nodes in the network. The team, which has a strong expertise in robustness for engineering problems, aims at contributing at the develpment of new robustness metrics in this biological context.

3.3. Stabilization of interconnected systems

- Linear systems: Analytic and algebraic approaches are considered for infinite-dimensional linear systems studied within the input-output framework.

  In the recent years, the Youla-Kučera parametrization (which gives the set of all stabilizing controllers of a system in terms of its coprime factorizations) has been the cornerstone of the success of the H_∞-control since this parametrization allows one to rewrite the problem of finding the optimal stabilizing controllers for a certain norm such as H_∞ or H_2 as affine, and thus, convex problem.

  A central issue studied in the team is the computation of such factorizations for a given infinite-dimensional linear system as well as establishing the links between stabilizability of a system for a certain norm and the existence of coprime factorizations for this system. These questions are fundamental for robust stabilization problems [1], [2], [8], [9].

  We also consider simultaneous stabilization since it plays an important role in the study of reliable stabilization, i.e. in the design of controllers which stabilize a finite family of plants describing a system during normal operating conditions and various failed modes (e.g. loss of sensors or actuators, changes in operating points) [9]. Moreover, we investigate strongly stabilizable systems [9], namely systems which can be stabilized by stable controllers, since they have a good ability to track reference inputs and, in practice, engineers are reluctant to use unstable controllers especially when the system is stable.

- Nonlinear systems

  The project aims at developing robust stabilization theory and methods for important classes of nonlinear systems that ensure good controller performance under uncertainty and time delays. The main techniques include techniques called backstepping and forwarding, contructions of strict Lyapunov functions through so-called "strictification" approaches [3] and construction of Lyapunov-Krasovskii functionals [4], [5], [6].

- Predictive control

  For highly complex systems described in the time-domain and which are submitted to constraints, predictive control seems to be well-adapted. This model based control method (MPC: Model Predictive Control) is founded on the determination of an optimal control sequence over a receding horizon. Due to its formulation in the time-domain, it is an effective tool for handling constraints and uncertainties which can be explicitly taken into account in the synthesis procedure [7]. The team considers how mutiparametric optimization can help to reduce the computational load of this method, allowing its effective use on real world constrained problems.

  The team also investigates stochastic optimization methods such as genetic algorithm, particle swarm optimization or ant colony [10] as they can be used to optimize any criterion and constraint whatever their mathematical structure is. The developed methodologies can be used by non specialists.

3.4. Synthesis of reduced complexity controllers

- PID controllers

  Even though the synthesis of control laws of a given complexity is not a new problem, it is still open, even for finite-dimensional linear systems. Our purpose is to search for good families of “simple” (e.g. low order) controllers for infinite-dimensional dynamical systems. Within our approach, PID candidates are first considered in the team [2], [87].

- Predictive control

  The synthesis of predictive control laws is concerned with the solution of multiparametric optimization problems. Reduced order controller constraints can be viewed as non convex constraints in the synthesis procedure. Such constraints can be taken into account with stochastic algorithms.

Finally, the development of algorithms based on both symbolic computation and numerical methods, and their implementations in dedicated Scilab/Matlab/Maple toolboxes are important issues in the project.

DOLPHIN Project-Team

3. Research Program

3.1. Hybrid multi-objective optimization methods

The success of metaheuristics is based on their ability to find efficient solutions in a reasonable time [58]. But with very large problems and/or multi-objective problems, efficiency of metaheuristics may be compromised. Hence, in this context it is necessary to integrate metaheuristics in more general schemes in order to develop even more efficient methods. For instance, this can be done by different strategies such as cooperation and parallelization.

The DOLPHIN project deals with *"a posteriori"* multi-objective optimization where the set of Pareto solutions (solutions of best compromise) have to be generated in order to give the decision maker the opportunity to choose the solution that interests him/her.

Population-based methods, such as evolutionary algorithms, are well fitted for multi-objective problems, as they work with a set of solutions [53], [57]. To be convinced one may refer to the list of references on Evolutionary Multi-objective Optimization maintained by Carlos A. Coello [0], which contains more than 5500 references. One of the objectives of the project is to propose advanced search mechanisms for intensification and diversification. These mechanisms have been designed in an adaptive manner, since their effectiveness is related to the landscape of the MOP and to the instance solved.

In order to assess the performances of the proposed mechanisms, we always proceed in two steps: first, we carry out experiments on academic problems, for which some best known results exist; second, we use real industrial problems to cope with large and complex MOPs. The lack of references in terms of optimal or best known Pareto set is a major problem. Therefore, the obtained results in this project and the test data sets will be available at the URL http://dolphin.lille.inria.fr/ at 'benchmark'.

3.1.1. Cooperation of metaheuristics

In order to benefit from the various advantages of the different metaheuristics, an interesting idea is to combine them. Indeed, the hybridization of metaheuristics allows the cooperation of methods having complementary behaviors. The efficiency and the robustness of such methods depend on the balance between the exploration of the whole search space and the exploitation of interesting areas.

Hybrid metaheuristics have received considerable interest these last years in the field of combinatorial optimization. A wide variety of hybrid approaches have been proposed in the literature and give very good results on numerous single objective optimization problems, which are either academic (traveling salesman problem, quadratic assignment problem, scheduling problem, etc) or real-world problems. This efficiency is generally due to the combinations of single-solution based methods (iterative local search, simulated annealing, tabu search, etc) with population-based methods (genetic algorithms, ants search, scatter search, etc). A taxonomy of hybridization mechanisms may be found in [62]. It proposes to decompose these mechanisms into four classes:

- *LRH class - Low-level Relay Hybrid*: This class contains algorithms in which a given metaheuristic is embedded into a single-solution metaheuristic. Few examples from the literature belong to this class.
- *LTH class - Low-level Teamwork Hybrid*: In this class, a metaheuristic is embedded into a population-based metaheuristic in order to exploit strengths of single-solution and population-based metaheuristics.

[0] http://www.lania.mx/~ccoello/EMOO/EMOObib.html

- *HRH class - High-level Relay Hybrid*: Here, self contained metaheuristics are executed in a sequence. For instance, a population-based metaheuristic is executed to locate interesting regions and then a local search is performed to exploit these regions.
- *HTH class - High-level Teamwork Hybrid*: This scheme involves several self-contained algorithms performing a search in parallel and cooperating. An example will be the island model, based on GAs, where the population is partitioned into small subpopulations and a GA is executed per subpopulation. Some individuals can migrate between subpopulations.

Let us notice that, hybrid methods have been studied in the mono-criterion case, their application in the multi-objective context is not yet widely spread. The objective of the DOLPHIN project is to integrate specificities of multi-objective optimization into the definition of hybrid models.

3.1.2. Cooperation between metaheuristics and exact methods

Until now only few exact methods have been proposed to solve multi-objective problems. They are based either on a Branch-and-bound approach, on the algorithm $A^{\star}$, or on dynamic programming. However, these methods are limited to two objectives and, most of the time, cannot be used on a complete large scale problem. Therefore, sub search spaces have to be defined in order to use exact methods. Hence, in the same manner as hybridization of metaheuristics, the cooperation of metaheuristics and exact methods is also a main issue in this project. Indeed, it allows us to use the exploration capacity of metaheuristics, as well as the intensification ability of exact methods, which are able to find optimal solutions in a restricted search space. Sub search spaces have to be defined along the search. Such strategies can be found in the literature, but they are only applied to mono-objective academic problems.

We have extended the previous taxonomy for hybrid metaheuristics to the cooperation between exact methods and metaheuristics. Using this taxonomy, we are investigating cooperative multi-objective methods. In this context, several types of cooperations may be considered, according to the way the metaheuristic and the exact method cooperate. For instance, a metaheuristic can use an exact method for intensification or an exact method can use a metaheuristic to reduce the search space.

Moreover, a part of the DOLPHIN project deals with studying exact methods in the multi-objective context in order: i) to be able to solve small size problems and to validate proposed heuristic approaches; ii) to have more efficient/dedicated exact methods that can be hybridized with metaheuristics. In this context, the use of parallelism will push back limits of exact methods, which will be able to explore larger size search spaces [55].

3.1.3. Goals

Based on the previous works on multi-objective optimization, it appears that to improve metaheuristics, it becomes essential to integrate knowledge about the problem structure. This knowledge can be gained during the search. This would allow us to adapt operators which may be specific for multi-objective optimization or not. The goal here is to design auto-adaptive methods that are able to react to the problem structure. Moreover, regarding the hybridization and the cooperation aspects, the objectives of the DOLPHIN project are to deepen these studies as follows:

- *Design of metaheuristics for the multi-objective optimization:* To improve metaheuristics, it becomes essential to integrate knowledge about the problem structure, which we may get during the execution. This would allow us to adapt operators that may be specific for multi-objective optimization or not. The goal here is to design auto-adaptive methods that are able to react to the problem structure.
- *Design of cooperative metaheuristics:* Previous studies show the interest of hybridization for a global optimization and the importance of problem structure study for the design of efficient methods. It is now necessary to generalize hybridization of metaheuristics and to propose adaptive hybrid models that may evolve during the search while selecting the appropriate metaheuristic. Multi-objective aspects have to be introduced in order to cope with the specificities of multi-objective optimization.

- *Design of cooperative schemes between exact methods and metaheuristics:* Once the study on possible cooperation schemes is achieved, we will have to test and compare them in the multi-objective context.
- *Design and conception of parallel metaheuristics:* Our previous works on parallel metaheuristics allow us to speed up the resolution of large scale problems. It could be also interesting to study the robustness of the different parallel models (in particular in the multi-objective case) and to propose rules that determine, given a specific problem, which kind of parallelism to use. Of course these goals are not disjoined and it will be interesting to simultaneously use hybrid metaheuristics and exact methods. Moreover, those advanced mechanisms may require the use of parallel and distributed computing in order to easily make cooperating methods evolve simultaneously and to speed up the resolution of large scale problems.
- *Validation:* In order to validate the obtained results we always proceed in two phases: validation on academic problems, for which some best known results exist and use on real problems (industrial) to cope with problem size constraints.

  Moreover, those advanced mechanisms are to be used in order to integrate the distributed multi-objective aspects in the ParadisEO platform (see the paragraph on software platform).

3.2. Parallel multi-objective optimization: models and software frameworks

Parallel and distributed computing may be considered as a tool to speedup the search to solve large MOPs and to improve the robustness of a given method. Moreover, the joint use of parallelism and cooperation allows improvements on the quality of the obtained Pareto sets. Following this objective, we will design and implement parallel models for metaheuristics (evolutionary algorithms, tabu search approach) and exact methods (branch-and-bound algorithm, branch-and-cut algorithm) to solve different large MOPs.

One of the goals of the DOLPHIN project is to integrate the developed parallel models into software frameworks. Several frameworks for parallel distributed metaheuristics have been proposed in the literature. Most of them focus only either on evolutionary algorithms or on local search methods. Only few frameworks are dedicated to the design of both families of methods. On the other hand, existing optimization frameworks either do not provide parallelism at all or just supply at most one parallel model. In this project, a new framework for parallel hybrid metaheuristics is proposed, named *Parallel and Distributed Evolving Objects (ParadisEO)* based on EO. The framework provides in a transparent way the hybridization mechanisms presented in the previous section, and the parallel models described in the next section. Concerning the developed parallel exact methods for MOPs, we will integrate them into well-known frameworks such as COIN.

3.2.1. Parallel models

According to the family of addressed metaheuristics, we may distinguish two categories of parallel models: parallel models that manage a single solution, and parallel models that handle a population of solutions. The major single solution-based parallel models are the following: the *parallel neighborhood exploration model* and the *multi-start model*.

- *The parallel neighborhood exploration model* is basically a "low level" model that splits the neighborhood into partitions that are explored and evaluated in parallel. This model is particularly interesting when the evaluation of each solution is costly and/or when the size of the neighborhood is large. It has been successfully applied to the mobile network design problem (see Application section).
- *The multi-start model* consists in executing in parallel several local searches (that may be heterogeneous), without any information exchange. This model raises particularly the following question: is it equivalent to execute k local searches during a time t than executing a single local search during $k \times t$? To answer this question we tested a multi-start Tabu search on the quadratic assignment problem. The experiments have shown that the answer is often landscape-dependent. For example, the multi-start model may be well-suited for landscapes with multiple basins.

Parallel models that handle a population of solutions are mainly: the *island model*, the *central model* and *the distributed evaluation of a single solution*. Let us notice that the last model may also be used with single-solution metaheuristics.

- In *the island model*, the population is split into several sub-populations distributed among different processors. Each processor is responsible of the evolution of one sub-population. It executes all the steps of the metaheuristic from the selection to the replacement. After a given number of generations (synchronous communication), or when a convergence threshold is reached (asynchronous communication), the migration process is activated. Then, exchanges of solutions between sub-populations are realized, and received solutions are integrated into the local sub-population.
- *The central (Master/Worker) model* allows us to keep the sequentiality of the original algorithm. The master centralizes the population and manages the selection and the replacement steps. It sends sub-populations to the workers that execute the recombination and evaluation steps. The latter returns back newly evaluated solutions to the master. This approach is efficient when the generation and evaluation of new solutions is costly.
- *The distributed evaluation model* consists in a parallel evaluation of each solution. This model has to be used when, for example, the evaluation of a solution requires access to very large databases (data mining applications) that may be distributed over several processors. It may also be useful in a multi-objective context, where several objectives have to be computed simultaneously for a single solution.

As these models have now been identified, our objective is to study them in the multi-objective context in order to use them advisedly. Moreover, these models may be merged to combine different levels of parallelism and to obtain more efficient methods [56], [61].

3.2.2. Goals

Our objectives focus on these issues are the following:

- *Design of parallel models for metaheuristics and exact methods for MOPs*: We will develop parallel cooperative metaheuristics (evolutionary algorithms and local search algorithms such as the Tabu search) for solving different large MOPs. Moreover, we are designing a new exact method, named PPM (Parallel Partition Method), based on branch and bound and branch and cut algorithms. Finally, some parallel cooperation schemes between metaheuristics and exact algorithms have to be used to solve MOPs in an efficient manner.
- *Integration of the parallel models into software frameworks*: The parallel models for metaheuristics will be integrated in the ParadisEO software framework. The proposed multi-objective exact methods must be first integrated into standard frameworks for exact methods such as COIN and BOB++. A *coupling* with ParadisEO is then needed to provide hybridization between metaheuristics and exact methods.
- *Efficient deployment of the parallel models on different parallel and distributed architectures including GRIDs*: The designed algorithms and frameworks will be efficiently deployed on non-dedicated networks of workstations, dedicated cluster of workstations and SMP (Symmetric Multi-processors) machines. For GRID computing platforms, peer to peer (P2P) middlewares (XtremWeb-Condor) will be used to implement our frameworks. For this purpose, the different optimization algorithms may be re-visited for their efficient deployment.

ECUADOR Project-Team

3. Research Program

3.1. Algorithmic Differentiation

Participants: Laurent Hascoët, Valérie Pascual, Ala Taftaf.

algorithmic differentiation (AD, aka Automatic Differentiation) Transformation of a program, that returns a new program that computes derivatives of the initial program, i.e. some combination of the partial derivatives of the program's outputs with respect to its inputs.

adjoint Mathematical manipulation of the Partial Derivative Equations that define a problem, obtaining new differential equations that define the gradient of the original problem's solution.

Algorithmic Differentiation (AD) differentiates *programs*. The input of AD is a source program P that, given some $X \in \mathbb{R}^n$, returns some $Y = F(X) \in \mathbb{R}^m$, for a differentiable F. AD generates a new source program P' that, given X, computes some derivatives of F [14].

The resulting P' reuses the control of P. For any given control, P is equivalent to a sequence of instructions, which is identified with a composition of vector functions. Thus, if

$$\begin{array}{lll} P & \text{is} & \{I_1; I_2; \cdots I_p;\}, \\ F & \text{then is} & f_p \circ f_{p-1} \circ \cdots \circ f_1, \end{array} \tag{4}$$

where each f_k is the elementary function implemented by instruction I_k. AD applies the chain rule to obtain derivatives of F. Calling X_k the values of all variables after instruction I_k, i.e. $X_0 = X$ and $X_k = f_k(X_{k-1})$, the Jacobian of F is

$$F'(X) = f'_p(X_{p-1}) \,.\, f'_{p-1}(X_{p-2}) \,.\, \cdots \,.\, f'_1(X_0) \tag{5}$$

which can be mechanically written as a sequence of instructions I'_k. Combining the I'_k with the control of P yields P', and therefore this differentiation is piecewise.

AD can be generalized to higher level derivatives, Taylor series, etc. In practice, many applications only need cheaper projections of $F'(X)$ such as:

- **Sensitivities**, defined for a given direction $\dot{X}$ in the input space as:

$$F'(X).\dot{X} = f'_p(X_{p-1}) \,.\, f'_{p-1}(X_{p-2}) \,.\, \cdots \,.\, f'_1(X_0) \,.\, \dot{X} \quad . \tag{6}$$

  This expression is easily computed from right to left, interleaved with the original program instructions. This is the *tangent mode* of AD.

- **Adjoints**, defined after transposition (F'^*), for a given weighting $\overline{Y}$ of the outputs as:

$$F'^*(X).\overline{Y} = f'^*_1(X_0).f'^*_2(X_1).\ \cdots\ .f'^*_{p-1}(X_{p-2}).f'^*_p(X_{p-1}).\overline{Y} \quad . \tag{7}$$

  This expression is most efficiently computed from right to left, because matrix×vector products are cheaper than matrix×matrix products. This defines the *adjoint mode* of AD, most effective for optimization, data assimilation [28], adjoint problems [23], or inverse problems.

Adjoint-mode AD turns out to make a very efficient program, at least theoretically [25]. The computation time required for the gradient is only a small multiple of the run-time of P. It is independent from the number of parameters n. In contrast, computing the same gradient with the *tangent mode* would require running the tangent differentiated program n times.

However, the X_k are required in the *inverse* of their computation order. If the original program *overwrites* a part of X_k, the differentiated program must restore X_k before it is used by $f'^*_{k+1}(X_k)$. Therefore, the central research problem of adjoint-mode AD is to make the X_k available in reverse order at the cheapest cost, using strategies that combine storage, repeated forward computation from available previous values, or even inverted computation from available later values.

Another research issue is to make the AD model cope with the constant evolution of modern language constructs. From the old days of Fortran77, novelties include pointers and dynamic allocation, modularity, structured data types, objects, vectorial notation and parallel communication. We keep developing our models and tools to handle these new constructs.

3.2. Static Analysis and Transformation of programs

Participants: Laurent Hascoët, Valérie Pascual, Ala Taftaf.

abstract syntax tree Tree representation of a computer program, that keeps only the semantically significant information and abstracts away syntactic sugar such as indentation, parentheses, or separators.

control flow graph Representation of a procedure body as a directed graph, whose nodes, known as basic blocks, each contain a sequence of instructions and whose arrows represent all possible control jumps that can occur at run-time.

abstract interpretation Model that describes program static analysis as a special sort of execution, in which all branches of control switches are taken concurrently, and where computed values are replaced by abstract values from a given *semantic domain*. Each particular analysis gives birth to a specific semantic domain.

data flow analysis Program analysis that studies how a given property of variables evolves with execution of the program. Data Flow analysis is static, therefore studying all possible run-time behaviors and making conservative approximations. A typical data-flow analysis is to detect, at any location in the source program, whether a variable is initialized or not.

data dependence analysis Program analysis that studies the itinerary of values during program execution, from the place where a value is defined to the places where it is used, and finally to the place where it is overwritten. The collection of all these itineraries is stored as *Def-Use and Use-Def chains* or as a *data dependence graph*, and data flow analysis most often rely on this graph.

data dependence graph Directed graph that relates accesses to program variables, from the write access that defines a new value to the read accesses that use this value, and from the read accesses to the write access that overwrites this value. Dependences express a partial order between operations, that must be preserved to preserve the program's result.

The most obvious example of a program transformation tool is certainly a compiler. Other examples are program translators, that go from one language or formalism to another, or optimizers, that transform a program to make it run better. AD is just one such transformation. These tools use sophisticated analysis [15]. These tools share their technological basis. More importantly, there are common mathematical models to specify and analyze them.

An important principle is *abstraction*: the core of a compiler should not bother about syntactic details of the compiled program. The optimization and code generation phases must be independent from the particular input programming language. This is generally achieved using language-specific *front-ends* and *back-ends*. Abstraction can go further: the internal representation becomes more language independent, and semantic constructs can be unified. Analysis can then concentrate on the semantics of a small set of constructs. We advocate an internal representation composed of three levels.

- At the top level is the *call graph*, whose nodes are modules and procedures. Arrows relate nodes that call or import one another. Recursion leads to cycles.
- At the middle level is the *flow graph*, one per procedure. It captures the control flow between atomic instructions. Loops lead to cycles.
- At the lowest level are abstract *syntax trees* for the individual atomic instructions. Semantic transformations can benefit from the representation of expressions as directed acyclic graphs, sharing common sub-expressions.

To each level belong symbol tables, nested to capture scoping.

Static program analysis can be defined on this internal representation, which is largely language independent. The simplest analyses on trees can be specified with inference rules [18], [26], [16]. But many analyses are more complex, and better defined on graphs than on trees. This is the case for *data-flow analyses*, that look for run-time properties of variables. Since flow graphs may be cyclic, these global analyses generally require an iterative resolution. *Data flow equations* is a practical formalism to describe data-flow analyses. Another formalism is described in [19], which is more precise because it can distinguish separate *instances* of instructions. However it is still based on trees, and its cost forbids application to large codes. *Abstract Interpretation* [20] is a theoretical framework to study complexity and termination of these analyses.

Data flow analyses must be carefully designed to avoid or control combinatorial explosion. At the call graph level, they can run bottom-up or top-down, and they yield more accurate results when they take into account the different call sites of each procedure, which is called *context sensitivity*. At the flow graph level, they can run forwards or backwards, and yield more accurate results when they take into account only the possible execution flows resulting from possible control, which is called *flow sensitivity*.

Even then, data flow analyses are limited, because they are static and thus have very little knowledge of actual run-time values. Far before reaching the very theoretical limit of *undecidability*, one reaches practical limitations to how much information one can infer from programs that use arrays [32], [21] or pointers. In general, conservative *over-approximations* are always made that lead to derivative code that is less efficient than possibly achievable.

3.3. Algorithmic Differentiation and Scientific Computing

Participants: Alain Dervieux, Laurent Hascoët, Bruno Koobus.

linearization In Scientific Computing, the mathematical model often consists of Partial Derivative Equations, that are discretized and then solved by a computer program. Linearization of these equations, or alternatively linearization of the computer program, predict the behavior of the model when small perturbations are applied. This is useful when the perturbations are effectively small, as in acoustics, or when one wants the sensitivity of the system with respect to one parameter, as in optimization.

adjoint state Consider a system of Partial Derivative Equations that define some characteristics of a system with respect to some input parameters. Consider one particular scalar characteristic. Its sensitivity, (or gradient) with respect to the input parameters can be defined as the solution of "adjoint" equations, deduced from the original equations through linearization and transposition. The solution of the adjoint equations is known as the adjoint state.

Scientific Computing provides reliable simulations of complex systems. For example it is possible to simulate the steady or unsteady 3D air flow around a plane that captures the physical phenomena of shocks and turbulence. Next comes optimization, one degree higher in complexity because it repeatedly simulates and applies optimization steps until an optimum is reached. We focus on gradient-based optimization.

We investigate several approaches to obtain the gradient, between two extremes:

- One can write an *adjoint system* of mathematical equations, then discretize it and program it by hand. This is mathematically sound [23], but very costly in development time. It also does not produce an exact gradient of the discrete function, and this can be a problem if using optimization methods based on descent directions.
- One can apply adjoint-mode AD (*cf*3.1) on the program that discretizes and solves the direct system. This gives in fact the adjoint of the discrete function computed by the program. Theoretical results [22] guarantee convergence of these derivatives when the direct program converges. This approach is highly mechanizable, but leads to massive use of storage and may require code transformation by hand [27], [30] to reduce memory usage.

If for instance the model is steady, or more generally when the computation uses a Fixed-Point iteration, tradeoffs exist between these two extremes [24], [17] that combine low storage consumption with possible automated adjoint generation. We advocate incorporating them into the AD model and into the AD tools.

GAMMA3 Project-Team (section vide)

GECO Project-Team

3. Research Program

3.1. Geometric control theory

The main research topic of the project-team will be **geometric control**, with a special focus on **control design**. The application areas that we target are control of quantum mechanical systems, neurogeometry and switched systems.

Geometric control theory provides a viewpoint and several tools, issued in particular from differential geometry, to tackle typical questions arising in the control framework: controllability, observability, stabilization, optimal control... [32], [66] The geometric control approach is particularly well suited for systems involving nonlinear and nonholonomic phenomena. We recall that nonholonomicity refers to the property of a velocity constraint that is not equivalent to a state constraint.

The expression **control design** refers here to all phases of the construction of a control law, in a mainly open-loop perspective: modeling, controllability analysis, output tracking, motion planning, simultaneous control algorithms, tracking algorithms, performance comparisons for control and tracking algorithms, simulation and implementation.

We recall that

- **controllability** denotes the property of a system for which any two states can be connected by a trajectory corresponding to an admissible control law ;
- **output tracking** refers to a control strategy aiming at keeping the value of some functions of the state arbitrarily close to a prescribed time-dependent profile. A typical example is **configuration tracking** for a mechanical system, in which the controls act as forces and one prescribes the position variables along the trajectory, while the evolution of the momenta is free. One can think for instance at the lateral movement of a car-like vehicle: even if such a movement is unfeasible, it can be tracked with arbitrary precision by applying a suitable control strategy;
- **motion planning** is the expression usually denoting the algorithmic strategy for selecting one control law steering the system from a given initial state to an attainable final one;
- **simultaneous control** concerns algorithms that aim at driving the system from two different initial conditions, with the same control law and over the same time interval, towards two given final states (one can think, for instance, at some control action on a fluid whose goal is to steer simultaneously two floating bodies.) Clearly, the study of which pairs (or n-uples) of states can be simultaneously connected thanks to an admissible control requires an additional controllability analysis with respect to the plain controllability mentioned above.

At the core of control design is then the notion of motion planning. Among the motion planning methods, a preeminent role is played by those based on the Lie algebra associated with the control system ([86], [73], [79]), those exploiting the possible flatness of the system ([60]) and those based on the continuation method ([98]). Optimal control is clearly another method for choosing a control law connecting two states, although it generally introduces new computational and theoretical difficulties.

Control systems with special structure, which are very important for applications are those for which the controls appear linearly. When the controls are not bounded, this means that the admissible velocities form a distribution in the tangent bundle to the state manifold. If the distribution is equipped with a smoothly varying norm (representing a cost of the control), the resulting geometrical structure is called *sub-Riemannian*. Sub-Riemannian geometry thus appears as the underlying geometry of the nonholonomic control systems, playing the same role as Euclidean geometry for linear systems. As such, its study is fundamental for control design. Moreover its importance goes far beyond control theory and is an active field of research both in differential geometry ([85]), geometric measure theory ([61], [36]) and hypoelliptic operator theory ([48]).

Other important classes of control systems are those modeling mechanical systems. The dynamics are naturally defined on the tangent or cotangent bundle of the configuration manifold, they have Lagrangian or Hamiltonian structure, and the controls act as forces. When the controls appear linearly, the resulting model can be seen somehow as a second-order sub-Riemannian structure (see [53]).

The control design topics presented above naturally extend to the case of distributed parameter control systems. The geometric approach to control systems governed by partial differential equations is a novel subject with great potential. It could complement purely analytical and numerical approaches, thanks to its more dynamical, qualitative and intrinsic point of view. An interesting example of this approach is the paper [33] about the controllability of Navier–Stokes equation by low forcing modes.

GEOSTAT Project-Team

3. Research Program

3.1. Multiscale description in terms of multiplicative cascade

GEOSTAT is studying complex signals under the point of view of *nonlinear* methods, in the sense of *nonlinear physics* i.e. the methodologies developed to study complex systems, with a strong emphasis on multiresolution analysis. Linear methods in signal processing refer to the standard point of view under which operators are expressed by simple convolutions with impulse responses. Linear methods in signal processing are widely used, from least-square deconvolution methods in adaptive optics to source-filter models in speech processing. Because of the absence of localization of the Fourier transform, linear methods are not successful to unlock the multiscale structures and cascading properties of variables which are of primary importance as stated by the physics of the phenomena. This is the reason why new approaches, such as DFA (Detrented Fluctuation Analysis), Time-frequency analysis, variations on curvelets [56] etc. have appeared during the last decades. Recent advances in dimensionality reduction, and notably in Compressive Sensing, go beyond the Nyquist rate in sampling theory using nonlinear reconstruction, but data reduction occur at random places, independently of geometric localization of information content, which can be very useful for acquisition purposes, but of lower impact in signal analysis. One important result obtained in GEOSTAT is the effective use of multiresolution analysis associated to optimal inference along the scales of a complex system. The multiresolution analysis is performed on dimensionless quantities given by the *singularity exponents* which encode properly the geometrical structures associated to multiscale organization. This is applied successfully in the derivation of high resolution ocean dynamics, or the high resolution mapping of gaseous exchanges between the ocean and the atmosphere; the latter is of primary importance for a quantitative evaluation of global warming. Understanding the dynamics of complex systems is recognized as a new discipline, which makes use of theoretical and methodological foundations coming from nonlinear physics, the study of dynamical systems and many aspects of computer science. One of the challenges is related to the question of *emergence* in complex systems: large-scale effects measurable macroscopically from a system made of huge numbers of interactive agents [48], [45], [61], [52]. Some quantities related to nonlinearity, such as Lyapunov exponents, Kolmogorov-Sinai entropy etc. can be computed at least in the phase space [46]. Consequently, knowledge from acquisitions of complex systems (which include *complex signals*) could be obtained from information about the phase space. A result from F. Takens [57] about strange attractors in turbulence has motivated the determination of discrete dynamical systems associated to time series [50], and consequently the theoretical determination of nonlinear characteristics associated to complex acquisitions. Emergence phenomena can also be traced inside complex signals themselves, by trying to localize information content geometrically. Fundamentally, in the nonlinear analysis of complex signals there are broadly two approaches: characterization by attractors (embedding and bifurcation) and time-frequency, multiscale/multiresolution approaches. Time-frequency analysis [47] and multiscale/multiresolution are the subjects of intense research and are profoundly reshaping the analysis of complex signals by nonlinear approaches [44], [49]. In real situations, the phase space associated to the acquisition of a complex phenomenon is unknown. It is however possible to relate, inside the signal's domain, local predictability to local reconstruction and deduce from that singularity exponents (SEs) [10] [6]. The SEs are defined at any point in the signal's domain, they relate, but are different, to other kinds of exponents used in the nonlinear analysis of complex signals. We are working on their relation with:

- properties in universality classses,
- the geometric localization of multiscale properties in complex signals,
- cascading characteristics of physical variables,
- optimal wavelets and inference in multiresolution analysis.

The alternative approach taken in GEOSTAT is microscopical, or geometrical: the multiscale structures which have their "fingerprint" in complex signals are being isolated in a single realization of the complex system, i.e. using the data of the signal itself, as opposed to the consideration of grand ensembles or a wide set of realizations. This is much harder than the ergodic approaches, but it is possible because a reconstruction formula such as the one derived in [58] is local and reconstruction in the signal's domain is related to predictability. This approach is analogous to the consideration of "microcanonical ensembles" in statistical mechanics.

Nonlinear signal processing is making use of quantities related to predictability. For instance the first Lyapunov exponent λ_1 is related, from Osedelec's theorem, to the limiting behaviour of the response, after a time t, to perturbation in the phase space $\log R_\tau(t)$:

$$\lambda_1 = \lim_{t\to\infty} \frac{1}{t} \langle \log R_\tau(t) \rangle \tag{8}$$

with $\langle \cdot \rangle$ being time average and R_τ the response to a perturbation [46]. In GEOSTAT our aim is to relate such classical quantities (among others) to the behaviour of SEs, which are defined by a limiting behaviour

$$\mu\left(\mathcal{B}_r(\mathbf{x})\right) \;=\; \alpha(\mathbf{x})\, r^{d+h(\mathbf{x})} \;+\; o\left(r^{d+h(\mathbf{x})}\right) \;\; (r \to 0) \tag{9}$$

(d: dimension of the signal's domain, μ: multiscale measure, typically whose density is the gradient's norm, $\mathcal{B}_r(\mathbf{x})$: ball of radius r centered at $\mathbf{x}$). For precise computation, SEs can be smoothly interpolated by projecting wavelets:

$$\mathcal{T}_\Psi \mu(\mathbf{x}, r) = \int_{\mathbb{R}^d} \mathrm{d}\mu(\mathbf{x}') \frac{1}{r^d} \Psi\left(\frac{\mathbf{x}-\mathbf{x}'}{r}\right) \tag{10}$$

(Ψ: mother wavelet, admissible or not), but the best numerical method in computing singularity exponents lies in the definition of a measure related to predictability [16]:

$$h(\mathbf{x}) = \frac{\log \mathcal{T}_\Psi \mu(\mathbf{x}, r_0) / \langle \mathcal{T}_\Psi \mu(\cdot, r_0) \rangle}{\log r_0} + o(\frac{1}{\log r_0}) \tag{11}$$

with: r_0 is a scale choosen to diminish the amplitude of the correction term, and $\langle \mathcal{T}_\Psi \mu(\cdot, r_0) \rangle$ is the average value of the wavelet projection (mother wavelet Ψ) over the whole signal. Singularity exponents computed with this formula generalize the elementary "gradient's norm" in a very statistically coherent way across the scales.

SEs are related to the framework of reconstructible systems, and consequently to predictability. They unlock the geometric localization of a multiscale structure in a complex signal:

$$\mathcal{F}_h = \{\mathbf{x} \in \Omega \;\mid\; h(\mathbf{x}) = h\}, \tag{12}$$

(Ω: signal's domain). This multiscale structure is a fundamental feature of a complex system. Indeed, let us take the explicit example of a signal which is an acquisition of a 3D turbulent fluid. The velocity field of the flow, $\mathbf{v}(\mathbf{x},t)$, is a solution of the Navier-Stokes equations. Fully Developped Turbulence (FDT) is defined as the regime observed when the Reynolds number $R \to \infty$, R being defined as the ratio of "viscous diffusion time" by "circulation time": $R = \frac{LV}{\nu}$,L and V being respectively characteristic length and velocity of the flow. The phase space of the associated dynamical system is infinite dimensional, while the dynamics of the flow possess one or more finite dimensional attractors. In the case of FDT, particles of the fluid in the continuum which are trapped around KAM invariant manifolds undergo random perturbations in their motion which accounts for the "boost" observed in turbulent diffusion. From there comes the observed behaviour for the energy spectrum (the law $\mathcal{E}(\mathbf{k}) \sim |\mathbf{k}|^{-5/3}$ within the inertial range), an observation that was the starting point of the Kolmogorov K41 theory, but is still not directly mathematically related from the Navier-Stokes equations. Intermittency is observed within the inertial range and is related to the fact that, in the case of FDT, symmetry is restored only in a statistical sense, a fact that has consequences on the quality of any nonlinear signal representation by frames or dictionaries.

The example of FDT as a standard "template" for developing general methods that apply to a vast class of complex systems and signals is of fundamental interest because, in FDT, the existence of a multiscale hierarchy (i.e. the collection of sets $\mathcal{F}_h$ of equation 5) which is of multifractal nature and geometrically localized can be derived from physical considerations. This geometric hierarchy of sets is responsible for the shape of the computed singularity spectra, which in turn is related to the statistical organization of information content in a signal. It explains scale invariance, a characteristic feature of complex signals. The analogy from statistical physics comes from the fact that singularity exponents are direct generalizations of *critical exponents* which explain the macroscopic properties of a system around critical points, and the quantitative characterization of *universality classes*, which allow the definition of methods and algorithms that apply to general complex signals and systems, and not only turbulent signals: signals which belong to a same universality class share common statistical organization. In GEOSTAT, the approach to singularity exponents is done within a microcanonical setting, which can interestingly be compared with other approaches such that wavelet leaders, WTMM or DFA. During the past decades, classical approaches (here called "canonical" because they use the analogy taken from the consideration of "canonical ensembles" in statistical mechanics) permitted the development of a well-established analogy taken from thermodynamics in the analysis of complex signals: if $\mathcal{F}$ is the free energy, $\mathcal{T}$ the temperature measured in energy units, $\mathcal{U}$ the internal energy per volume unit $\mathcal{S}$ the entropy and $\widehat{\beta} = 1/\mathcal{T}$, then the scaling exponents associated to moments of intensive variables $p \to \tau_p$ corresponds to $\widehat{\beta}\mathcal{F}$, $\mathcal{U}(\widehat{\beta})$ corresponds to the singularity exponents values, and $\mathcal{S}(\mathcal{U})$ to the singularity spectrum.

The singularity exponents belong to a universality class, independently of microscopic properties in the phase space of various complex systems, and beyond the particular case of turbulent data (where the existence of a multiscale hierarchy, of multifractal nature, can be inferred directly from physical considerations). They describe common multiscale statistical organizations in different complex systems [55], and this is why GEOSTAT is working on nonlinear signal processing tools that are applied to very different types of signals. The methodological framework used in GEOSTAT for analyzing complex signals is different from, but related to, the "canonical" apparatus developed in recent years (WTMM method, wavelet leaders etc.). In the microcanonical approach developed, geometrically localized singularity exponents relate to a "microcanonical" description of multiplicative cascades observed in complex systems. Indeed, it can be shown that p-dissipation at scale r associated to a fixed interval $]p, p+\Delta p[$, $\epsilon_r^{(p,\Delta p)}$, behaves in the limit $\Delta p \to 0$ as

$$\epsilon_r^{(p)} = \lim_{\Delta p \to 0} \epsilon_r^{(p,\Delta p)} = \left(\epsilon_r^{(\infty)}\right)^{h(p)/h_\infty} \tag{13}$$

which indicates the existence of a relation between the multiscale hierarchy and the geometric localization of the cascade in complex systems.

The GEOSTAT team is working particularly on the very important subject of *optimal wavelets* which are wavelets ψ that "split" the signal projections between two different scales $\mathbf{r_1} < \mathbf{r_2}$ in such a way that there exists an injection term $\zeta_{r_1/r_2}(\mathbf{x})$, independent of the process $\mathcal{T}_\psi[\mathbf{s}](\mathbf{x},\mathbf{r})$ with:

$$\mathcal{T}_\psi[\mathbf{s}](\mathbf{x}, \mathbf{r_1}) = \zeta_{r_1/r_2}(\mathbf{x})\mathcal{T}_\psi[\mathbf{s}](\mathbf{x}, \mathbf{r_2}) \tag{14}$$

($\mathbf{r_1} < \mathbf{r_2}$: two scales of observation, ζ: injection variable between the scales, ψ: optimal wavelet). The **multiresolution analysis** associated to optimal wavelets is particularly interesting because it reflects, in an optimal way, the cross-scale information transfer in a complex system. These wavelets are related to persistence along the scales and lead to multiresolution analysis whose coefficients verify

$$\alpha_s = \eta_1 \alpha_f + \eta_2 \tag{15}$$

with α_s and α_f referring to child and parent coefficients, η_1 and η_2 are random variables independent of α_s and α_f and also independent of each other.

For example we give some insight about the collaboration with LEGOS Dynbio team [0] about high-resolution ocean dynamics from microcanonical formulations in nonlinear complex signal analysis. LPEs relate to the geometric structures linked with the cascading properties of indefinitely divisible variables in turbulent flows. Cascading properties can be represented by optimal wavelets (OWs); this opens new and fascinating directions of research for the determination of ocean motion field at high spatial resolution. OWs in a microcanonical sense pave the way for the determination of the energy injection mechanisms between the scales. From this results a new method for the complete evaluation of oceanic motion field; it consists in propagating along the scales the norm and the orientation of ocean dynamics deduced at low spatial resolution (geostrophic from altimetry and a part of ageostrophic from wind stress products). Using this approach, there is no need to use several temporal occurrences. Instead, the proper determination of the turbulent cascading and energy injection mechanisms in oceanographic signals allows the determination of oceanic motion field at the SST or Ocean colour spatial resolution (pixel size: 4 kms). We use the Regional Ocean Modelling System (ROMS) to validate the results on simulated data and compare the motion fields obtained with other techniques [17].

3.2. Excitable systems

Highly promising results are obtained in the application of nonlinear signal processing and multiscale techniques to the localization of heart fibrillation phenomenon acquired from a real patient and mapped over a reconstructed 3D surface of the heart. The notion of *source field*, defined in GEOSTAT from the computation of derivative measures related to the singularity exponents allows the localization of arrythmic phenomena inside the heart [7].

In speech analysis, we use the concept of the Most Singular Manifold (MSM) to localize critical events in domain of this signal. We show that in case of voiced speech signals, the MSM coincides with the instants of significant excitation of the vocal tract system. It is known that these major excitations occur when the glottis is closed, and hence, they are called the Glottal Closure Instants (GCI). We use the MSM to develop a reliable and noise robust GCI detection algorithm and we evaluate our algorithm using contemporaneous Electro-Glotto-Graph (EGG) recordings.

[0] http://www.legos.obs-mip.fr/recherches/equipes/dynbio.

I4S Project-Team

3. Research Program

3.1. Introduction

In this section, the main features for the key monitoring issues, namely identification, detection, and diagnostics, are provided, and a particular instantiation relevant for vibration monitoring is described.

It should be stressed that the foundations for identification, detection, and diagnostics, are fairly general, if not generic. Handling high order linear dynamical systems, in connection with finite elements models, which call for using subspace-based methods, is specific to vibration-based SHM. Actually, one particular feature of model-based sensor information data processing as exercised in I4S, is the combined use of black-box or semi-physical models together with physical ones. Black-box and semi-physical models are, for example, eigenstructure parameterizations of linear MIMO systems, of interest for modal analysis and vibration-based SHM. Such models are intended to be identifiable. However, due to the large model orders that need to be considered, the issue of model order selection is really a challenge. Traditional advanced techniques from statistics such as the various forms of Akaike criteria (AIC, BIC, MDL, ...) do not work at all. This gives rise to new research activities specific to handling high order models.

Our approach to monitoring assumes that a model of the monitored system is available. This is a reasonable assumption, especially within the SHM areas. The main feature of our monitoring method is its intrinsic ability to the early warning of small deviations of a system with respect to a reference (safe) behavior under usual operating conditions, namely without any artificial excitation or other external action. Such a normal behavior is summarized in a reference parameter vector θ_0, for example a collection of modes and mode-shapes.

3.2. Identification

The behavior of the monitored continuous system is assumed to be described by a parametric model $\{\mathbf{P}_\theta\,,\,\theta\in\Theta\}$, where the distribution of the observations $(Z_0,...,Z_N)$ is characterized by the parameter vector $\theta\in\Theta$. An *estimating function*, for example of the form :

$$\mathcal{K}_N(\theta)=1/N\sum_{k=0}^{N}K(\theta,Z_k)$$

is such that $\mathbf{E}_\theta[\mathcal{K}_N(\theta)]=0$ for all $\theta\in\Theta$. In many situations, $\mathcal{K}$ is the gradient of a function to be minimized : squared prediction error, log-likelihood (up to a sign), ... For performing model identification on the basis of observations $(Z_0,...,Z_N)$, an estimate of the unknown parameter is then [61] :

$$\widehat{\theta}_N=\arg\{\theta\in\Theta\,:\,\mathcal{K}_N(\theta)=0\}$$

In many applications, such an approach must be improved in the following directions :

- *Recursive estimation:* the ability to compute $\widehat{\theta}_{N+1}$ simply from $\widehat{\theta}_N$;
- *Adaptive estimation:* the ability to *track* the true parameter θ^* when it is time-varying.

3.3. Detection

Our approach to on-board detection is based on the so-called asymptotic statistical local approach, which we have extended and adapted [5], [4], [2]. It is worth noticing that these investigations of ours have been initially motivated by a vibration monitoring application example. It should also be stressed that, as opposite to many monitoring approaches, our method does not require repeated identification for each newly collected data sample.

For achieving the early detection of small deviations with respect to the normal behavior, our approach generates, on the basis of the reference parameter vector θ_0 and a new data record, indicators which automatically perform :

- The early detection of a slight mismatch between the model and the data;
- A preliminary diagnostics and localization of the deviation(s);
- The tradeoff between the magnitude of the detected changes and the uncertainty resulting from the estimation error in the reference model and the measurement noise level.

These indicators are computationally cheap, and thus can be embedded. This is of particular interest in some applications, such as flutter monitoring.

As in most fault detection approaches, the key issue is to design a *residual*, which is ideally close to zero under normal operation, and has low sensitivity to noises and other nuisance perturbations, but high sensitivity to small deviations, before they develop into events to be avoided (damages, faults, ...). The originality of our approach is to :

- *Design* the residual basically as a *parameter estimating function*,
- *Evaluate* the residual thanks to a kind of central limit theorem, stating that the residual is asymptotically Gaussian and reflects the presence of a deviation in the parameter vector through a change in its own mean vector, which switches from zero in the reference situation to a non-zero value.

This is actually a strong result, which transforms any detection problem concerning a parameterized stochastic *process* into the problem of monitoring the mean of a Gaussian *vector*.

The behavior of the monitored system is again assumed to be described by a parametric model $\{\mathbf{P}_\theta\ ,\ \theta \in \Theta\}$, and the safe behavior of the process is assumed to correspond to the parameter value θ_0. This parameter often results from a preliminary identification based on reference data, as in module 3.2 .

Given a new N-size sample of sensors data, the following question is addressed : *Does the new sample still correspond to the nominal model* $\mathbf{P}_{\theta_0}$ *?* One manner to address this generally difficult question is the following. The asymptotic local approach consists in deciding between the nominal hypothesis and a *close* alternative hypothesis, namely :

$$\text{(Safe) } \mathbf{H}_0 :\ \theta = \theta_0 \quad \text{ and } \quad \text{(Damaged) } \mathbf{H}_1 :\ \theta = \theta_0 + \eta/\sqrt{N} \tag{16}$$

where η is an unknown but fixed change vector. A residual is generated under the form :

$$\zeta_N = 1/\sqrt{N}\ \sum_{k=0}^{N} K(\theta_0, Z_k) = \sqrt{N}\ \mathcal{K}_N(\theta_0)\ . \tag{17}$$

If the matrix $\mathcal{J}_N = -\,\mathbf{E}_{\theta_0}[\partial \mathcal{K}_N(\theta_0)]$ converges towards a limit $\mathcal{J}$, then, under mild mixing and stationarity assumptions, the central limit theorem shows [60] that the residual is asymptotically Gaussian :

$$\zeta_N \xrightarrow[N \to \infty]{} \begin{cases} \mathcal{N}(0, \Sigma) & \text{under } \mathbf{P}_{\theta_0}\ , \\ \\ \mathcal{N}(\mathcal{J}\,\eta, \Sigma) & \text{under } \mathbf{P}_{\theta_0+\eta/\sqrt{N}}\ , \end{cases} \tag{18}$$

where the asymptotic covariance matrix Σ can be estimated, and manifests the deviation in the parameter vector by a change in its own mean value. Then, deciding between $\eta = 0$ and $\eta \neq 0$ amounts to compute the following χ^2-test, provided that $\mathcal{J}$ is full rank and Σ is invertible :

$$\chi^2 = \overline{\zeta}^{\,T}\, \mathbf{F}^{-1}\, \overline{\zeta} \gtrless \lambda\ . \tag{19}$$

where

$$\overline{\zeta} \triangleq \mathcal{J}^T \Sigma^{-1} \zeta_N \quad \text{and} \quad \mathbf{F} \triangleq \mathcal{J}^T \Sigma^{-1} \mathcal{J} \tag{20}$$

With this approach, it is possible to decide, with a quantifiable error level, if a residual value is significantly different from zero, for assessing whether a fault/damage has occurred. It should be stressed that the residual and the sensitivity and covariance matrices $\mathcal{J}$ and Σ can be evaluated (or estimated) for the nominal model. In particular, it is *not* necessary to re-identify the model, and the sensitivity and covariance matrices can be pre-computed off-line.

3.4. Diagnostics

A further monitoring step, often called *fault isolation*, consists in determining which (subsets of) components of the parameter vector θ have been affected by the change. Solutions for that are now described. How this relates to diagnostics is addressed afterwards.

The question: *which (subsets of) components of θ have changed ?*, can be addressed using either nuisance parameters elimination methods or a multiple hypotheses testing approach [59].

In most SHM applications, a complex physical system, characterized by a generally non identifiable parameter vector Φ has to be monitored using a simple (black-box) model characterized by an identifiable parameter vector θ. A typical example is the vibration monitoring problem for which complex finite elements models are often available but not identifiable, whereas the small number of existing sensors calls for identifying only simplified input-output (black-box) representations. In such a situation, two different diagnosis problems may arise, namely diagnosis in terms of the black-box parameter θ and diagnosis in terms of the parameter vector Φ of the underlying physical model.

The isolation methods sketched above are possible solutions to the former. Our approach to the latter diagnosis problem is basically a detection approach again, and not a (generally ill-posed) inverse problem estimation approach [3]. The basic idea is to note that the physical sensitivity matrix writes $\mathcal{J}\mathcal{J}_{\Phi\theta}$, where $\mathcal{J}_{\Phi\theta}$ is the Jacobian matrix at Φ_0 of the application $\Phi \mapsto \theta(\Phi)$, and to use the sensitivity test for the components of the parameter vector Φ. Typically this results in the following type of directional test :

$$\chi^2_\Phi = \zeta^T \Sigma^{-1} \mathcal{J}\mathcal{J}_{\Phi\theta} \left(\mathcal{J}^T_{\Phi\theta} \mathcal{J}^T \Sigma^{-1} \mathcal{J}\mathcal{J}_{\Phi\theta}\right)^{-1} \mathcal{J}^T_{\Phi\theta} \mathcal{J}^T \Sigma^{-1} \zeta \gtrless \lambda \ . \tag{21}$$

It should be clear that the selection of a particular parameterization Φ for the physical model may have a non negligible influence on such type of tests, according to the numerical conditioning of the Jacobian matrices $\mathcal{J}_{\Phi\theta}$.

As a summary, the machinery in modules 3.2 , 3.3 and 3.4 provides us with a generic framework for designing monitoring algorithms for continuous structures, machines and processes. This approach assumes that a model of the monitored system is available. This is a reasonable assumption within the field of applications since most mechanical processes rely on physical principles which write in terms of equations, providing us with models. These important *modeling* and *parameterization* issues are among the questions we intend to investigate within our research program.

The key issue to be addressed within each parametric model class is the residual generation, or equivalently the choice of the *parameter estimating function*.

3.5. Subspace-based identification and detection

For reasons closely related to the vibrations monitoring applications, we have been investigating subspace-based methods, for both the identification and the monitoring of the eigenstructure (λ, ϕ_λ) of the state transition matrix F of a linear dynamical state-space system :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & F\ X_k + V_{k+1} \\ Y_k & = & H\ X_k \end{array} \right. , \tag{22}$$

namely the $(\lambda, \varphi_\lambda)$ defined by :

$$\det\ (F - \lambda\ I) = 0, \quad (F - \lambda\ I)\ \phi_\lambda = 0, \quad \varphi_\lambda \triangleq H\ \phi_\lambda \tag{23}$$

The (canonical) parameter vector in that case is :

$$\theta \triangleq \left(\begin{array}{c} \Lambda \\ \mathrm{vec}\Phi \end{array} \right) \tag{24}$$

where Λ is the vector whose elements are the eigenvalues λ, Φ is the matrix whose columns are the φ_λ's, and vec is the column stacking operator.

Subspace-based methods is the generic name for linear systems identification algorithms based on either time domain measurements or output covariance matrices, in which different subspaces of Gaussian random vectors play a key role [62]. A contribution of ours, minor but extremely fruitful, has been to write the output-only covariance-driven subspace identification method under a form that involves a parameter estimating function, from which we define a *residual adapted to vibration monitoring* [1]. This is explained next.

3.5.1. Covariance-driven subspace identification.

Let $R_i \triangleq \mathbf{E}\left(Y_k\ Y_{k-i}^T\right)$ and:

$$\mathcal{H}_{p+1,q} \triangleq \left(\begin{array}{cccc} R_0 & R_1 & \vdots & R_{q-1} \\ R_1 & R_2 & \vdots & R_q \\ \vdots & \vdots & \vdots & \vdots \\ R_p & R_{p+1} & \vdots & R_{p+q-1} \end{array} \right) \triangleq \mathrm{Hank}\left(R_i\right) \tag{25}$$

be the output covariance and Hankel matrices, respectively; and: $G \triangleq \mathbf{E}\left(X_k Y_k^T\right)$. Direct computations of the R_i's from the equations (10) lead to the well known key factorizations :

$$\begin{array}{rcl} R_i & = & HF^iG \\ \mathcal{H}_{p+1,q} & = & \mathcal{O}_{p+1}(H,F)\ \mathcal{C}_q(F,G) \end{array} \tag{26}$$

where:

$$\mathcal{O}_{p+1}(H,F) \triangleq \left(\begin{array}{c} H \\ HF \\ \vdots \\ HF^p \end{array} \right) \quad \text{and} \quad \mathcal{C}_q(F,G) \triangleq (G\ FG\ \cdots\ F^{q-1}G) \tag{27}$$

are the observability and controllability matrices, respectively. The observation matrix H is then found in the first block-row of the observability matrix $\mathcal{O}$. The state-transition matrix F is obtained from the shift invariance property of $\mathcal{O}$. The eigenstructure (λ, ϕ_λ) then results from (11).

Since the actual model order is generally not known, this procedure is run with increasing model orders.

3.5.2. Model parameter characterization.

Choosing the eigenvectors of matrix F as a basis for the state space of model (10) yields the following representation of the observability matrix:

$$\mathcal{O}_{p+1}(\theta) = \begin{pmatrix} \Phi \\ \Phi\Delta \\ \vdots \\ \Phi\Delta^p \end{pmatrix} \tag{28}$$

where $\Delta \triangleq \text{diag}(\Lambda)$, and Λ and Φ are as in (12). Whether a nominal parameter θ_0 fits a given output covariance sequence $(R_j)_j$ is characterized by [1]:

$$\mathcal{O}_{p+1}(\theta_0) \text{ and } \mathcal{H}_{p+1,q} \text{ have the same left kernel space.} \tag{29}$$

This property can be checked as follows. From the nominal θ_0, compute $\mathcal{O}_{p+1}(\theta_0)$ using (16), and perform e.g. a singular value decomposition (SVD) of $\mathcal{O}_{p+1}(\theta_0)$ for extracting a matrix U such that:

$$U^T\, U = I_s \text{ and } U^T\, \mathcal{O}_{p+1}(\theta_0) = 0 \tag{30}$$

Matrix U is not unique (two such matrices relate through a post-multiplication with an orthonormal matrix), but can be regarded as a function of θ_0. Then the characterization writes:

$$U(\theta_0)^T\, \mathcal{H}_{p+1,q} = 0 \tag{31}$$

3.5.3. Residual associated with subspace identification.

Assume now that *a reference θ_0 and a new sample $Y_1, \cdots, Y_N$ are available.* For checking whether the data agree with θ_0, the idea is to compute the empirical Hankel matrix $\widehat{\mathcal{H}}_{p+1,q}$:

$$\widehat{\mathcal{H}}_{p+1,q} \triangleq \text{Hank}\left(\widehat{R}_i\right), \quad \widehat{R}_i \triangleq 1/(N-i) \sum_{k=i+1}^{N} Y_k\, Y_{k-i}^T \tag{32}$$

and to define the residual vector:

$$\zeta_N(\theta_0) \triangleq \sqrt{N}\ \text{vec}\left(U(\theta_0)^T\, \widehat{\mathcal{H}}_{p+1,q}\right) \tag{33}$$

Let θ be the actual parameter value for the system which generated the new data sample, and $\mathbf{E}_\theta$ be the expectation when the actual system parameter is θ. From (19), we know that $\zeta_N(\theta_0)$ has zero mean when no change occurs in θ, and nonzero mean if a change occurs. Thus $\zeta_N(\theta_0)$ plays the role of a residual.

It is our experience that this residual has highly interesting properties, both for damage detection [1] and localization [3], and for flutter monitoring [8].

3.5.4. Other uses of the key factorizations.

Factorization (3.5.1) is the key for a characterization of the canonical parameter vector θ in (12), and for deriving the residual. Factorization (14) is also the key for :

- Proving consistency and robustness results [6];
- Designing an extension of covariance-driven subspace identification algorithm adapted to the presence and fusion of non-simultaneously recorded multiple sensors setups [7];
- Proving the consistency and robustness of this extension [9];
- Designing various forms of *input-output* covariance-driven subspace identification algorithms adapted to the presence of both known inputs and unknown excitations [10].

3.5.5. Research program

The research will first focus on the extension and implementation of current techniques as developed in I4S and IFSTTAR. Before doing any temperature rejection on large scale structures as planned, we need to develop good and accurate models of thermal fields. We also need to develop robust and efficient versions of our algorithms, mainly the subspace algorithms before envisioning linking them with physical models. Briefly, we need to mature our statistical toolset as well as our physical modeling before mixing them together later on.

3.5.5.1. Direct vibration modeling under temperature changes

This task builds upon what has been achieved in the CONSTRUCTIF project, where a simple formulation of the temperature effect has been exhibited, based on relatively simple assumptions. The next step is to generalize this modeling to a realistic large structure under complex thermal changes. Practically, temperature and resulting structural prestress and pre strains of thermal origin are not uniform and civil structures are complex. This leads to a fully 3D temperature field, not just a single value. Inertia effects also forbid a trivial prediction of the temperature based on current sensor outputs while ignoring past data. On the other side, the temperature is seen as a nuisance. That implies that any damage detection procedure has first to correct the temperature effect prior to any detection.

Modeling vibrations of structures under thermal prestress does and will play an important role in the static correction of kinematic measurements, in health monitoring methods based on vibration analysis as well as in durability and in the active or semi-active control of civil structures that by nature are operated under changing environmental conditions. As a matter of fact, using temperature and dynamic models the project aims at correcting the current vibration state from induced temperature effects, such that damage detection algorithms rely on a comparison of this thermally corrected current vibration state with a reference state computed or measured at a reference temperature. This approach is expected to cure damage detection algorithms from the environmental variations.

I4S will explore various ways of implementing this concept, notably within the FUI SIPRIS project.

3.5.5.2. Damage localization algorithms (in the case of localized damages such as cracks)

During the CONSTRUCTIF project, both feasibility and efficiency of some damage detection and localization algorithms were proved. Those methods are based on the tight coupling of statistical algorithms with finite element models. It has been shown that effective localization of some damaged elements was possible, and this was validated on a numerical simulated bridge deck model. Still, this approach has to be validated on real structures.

On the other side, new localization algorithms are currently investigated such as the one developed conjointly with University of Boston and tested within the framework of FP7 ISMS project. These algorithms will be implemented and tested on the PEGASE platform as well as all our toolset.

When possible, link with temperature rejection will be done along the lines of what has been achieved in the CONSTRUCTIF project.

3.5.5.3. Uncertainty quantification for system identification algorithms

Some emphasis will be put on expressing confidence intervals for system identification. It is a primary goal to take into account the uncertainty within the identification procedure, using either identification algorithms derivations or damage detection principles. Such algorithms are critical for both civil and aeronautical structures monitoring. It has been shown that confidence intervals for estimation parameters can theoretically be related to the damage detection techniques and should be computed as a function of the Fisher information matrix associated to the damage detection test. Based on those assumptions, it should be possible to obtain confidence intervals for a large class of estimates, from damping to finite elements models. Uncertainty considerations are also deeply investigated in collaboration with Dassault Aviation in Mellinger PhD thesis or with Northeastern University, Boston, within Gallegos PhD thesis.

3.5.5.4. Reflectometry-based methods for civil engineering structure health monitoring

For mechanical structures with a dominating geometrical axis so that they can be approximately considered one dimensional structures, some reflectometry-based methods initially developed for electrical cable monitoring have proved efficient for their health monitoring. Typical applications of such methods have been validated for the monitoring of external post-tensioned cables built with concrete bridges. Further studies are necessary to generalize this technology to other mechanical structures.

3.5.5.5. PEGASE platform

A new iteration called PEGASE 2 of our wireless platform has to be finalized (see Software section), in particular:

- Validation of PEGASE 2 mother board for its ability to recover energy from solar cells. Writing resulting abacus and user-guides...
- Discover and manage the DSP Library of PEGASE 2 (TI 5330 processor)
- Finalizing its main daughter boards:
  - 8 synchronous analog channel daughter board (finalized at 90
  - validation of the POE (Power Over Ethernet) daughter board
  - validation of the 3G daughter board (for GSM links)
  - Finalizing the supervisor (Matlab plugin...)

Figure 1. PEGASE board

IPSO Project-Team

3. Research Program

3.1. Structure-preserving numerical schemes for solving ordinary differential equations

Participants: François Castella, Philippe Chartier, Erwan Faou, Vilmart Gilles.

ordinary differential equation, numerical integrator, invariant, Hamiltonian system, reversible system, Lie-group system

In many physical situations, the time-evolution of certain quantities may be written as a Cauchy problem for a differential equation of the form

$$\begin{aligned} y'(t) &= f(y(t)), \\ y(0) &= y_0. \end{aligned} \tag{34}$$

For a given y_0, the solution $y(t)$ at time t is denoted $\varphi_t(y_0)$. For fixed t, φ_t becomes a function of y_0 called the *flow* of (1). From this point of view, a numerical scheme with step size h for solving (1) may be regarded as an approximation Φ_h of φ_h. One of the main questions of *geometric integration* is whether *intrinsic* properties of φ_t may be passed on to Φ_h.

This question can be more specifically addressed in the following situations:

3.1.1. Reversible ODEs

The system (1) is said to be ρ-reversible if there exists an involutive linear map ρ such that

$$\rho \circ \varphi_t = \varphi_t^{-1} \circ \rho = \varphi_{-t} \circ \rho. \tag{35}$$

It is then natural to require that Φ_h satisfies the same relation. If this is so, Φ_h is said to be *symmetric*. Symmetric methods for reversible systems of ODEs are just as much important as *symplectic* methods for Hamiltonian systems and offer an interesting alternative to symplectic methods.

3.1.2. ODEs with an invariant manifold

The system (1) is said to have an invariant manifold g whenever

$$\mathcal{M} = \{y \in \mathbb{R}^n; g(y) = 0\} \tag{36}$$

is kept *globally* invariant by φ_t. In terms of derivatives and for sufficiently differentiable functions f and g, this means that

$$\forall\, y \in \mathcal{M},\ g'(y)f(y) = 0.$$

As an example, we mention Lie-group equations, for which the manifold has an additional group structure. This could possibly be exploited for the space-discretisation. Numerical methods amenable to this sort of problems have been reviewed in a recent paper [60] and divided into two classes, according to whether they use g explicitly or through a projection step. In both cases, the numerical solution is forced to live on the manifold at the expense of some Newton's iterations.

3.1.3. Hamiltonian systems

Hamiltonian problems are ordinary differential equations of the form:

$$\begin{array}{rcccl} \dot{p}(t) & = & -\nabla_q H(p(t),q(t)) & \in & \mathbb{R}^d \\ \dot{q}(t) & = & \nabla_p H(p(t),q(t)) & \in & \mathbb{R}^d \end{array} \tag{37}$$

with some prescribed initial values $(p(0),q(0)) = (p_0,q_0)$ and for some scalar function H, called the Hamiltonian. In this situation, H is an invariant of the problem. The evolution equation (4) can thus be regarded as a differential equation on the manifold

$$\mathcal{M} = \{(p,q) \in \mathbb{R}^d \times \mathbb{R}^d; H(p,q) = H(p_0,q_0)\}.$$

Besides the Hamiltonian function, there might exist other invariants for such systems: when there exist d invariants in involution, the system (4) is said to be *integrable*. Consider now the parallelogram P originating from the point $(p,q) \in \mathbb{R}^{2d}$ and spanned by the two vectors $\xi \in \mathbb{R}^{2d}$ and $\eta \in \mathbb{R}^{2d}$, and let $\omega(\xi,\eta)$ be the sum of the *oriented* areas of the projections over the planes (p_i,q_i) of P,

$$\omega(\xi,\eta) = \xi^T J \eta,$$

where J is the *canonical symplectic* matrix

$$J = \begin{bmatrix} 0 & I_d \\ -I_d & 0 \end{bmatrix}.$$

A continuously differentiable map g from $\mathbb{R}^{2d}$ to itself is called symplectic if it preserves ω, i.e. if

$$\omega(g'(p,q)\xi, g'(p,q)\eta) = \omega(\xi,\eta).$$

A fundamental property of Hamiltonian systems is that their exact flow is symplectic. Integrable Hamiltonian systems behave in a very remarkable way: as a matter of fact, their invariants persist under small perturbations, as shown in the celebrated theory of Kolmogorov, Arnold and Moser. This behavior motivates the introduction of *symplectic* numerical flows that share most of the properties of the exact flow. For practical simulations of Hamiltonian systems, symplectic methods possess an important advantage: the error-growth as a function of time is indeed linear, whereas it would typically be quadratic for non-symplectic methods.

3.1.4. Differential-algebraic equations

Whenever the number of differential equations is insufficient to determine the solution of the system, it may become necessary to solve the differential part and the constraint part altogether. Systems of this sort are called differential-algebraic systems. They can be classified according to their index, yet for the purpose of this expository section, it is enough to present the so-called index-2 systems

$$\begin{array}{rcl} \dot{y}(t) & = & f(y(t),z(t)), \\ 0 & = & g(y(t)), \end{array} \tag{38}$$

where initial values $(y(0),z(0)) = (y_0,z_0)$ are given and assumed to be consistent with the constraint manifold. By constraint manifold, we imply the intersection of the manifold

$$\mathcal{M}_1 = \{y \in \mathbb{R}^n, g(y) = 0\}$$

and of the so-called hidden manifold

$$\mathcal{M}_2 = \{(y,z) \in \mathbb{R}^n \times \mathbb{R}^m, \frac{\partial g}{\partial y}(y) f(y,z) = 0\}.$$

This manifold $\mathcal{M} = \mathcal{M}_1 \bigcap \mathcal{M}_2$ is the manifold on which the exact solution $(y(t), z(t))$ of (5) lives.

There exists a whole set of schemes which provide a numerical approximation lying on $\mathcal{M}_1$. Furthermore, this solution can be projected on the manifold $\mathcal{M}$ by standard projection techniques. However, it it worth mentioning that a projection destroys the symmetry of the underlying scheme, so that the construction of a symmetric numerical scheme preserving $\mathcal{M}$ requires a more sophisticated approach.

3.2. Highly-oscillatory systems

Participants: François Castella, Philippe Chartier, Nicolas Crouseilles, Erwan Faou, Florian Méhats, Mohammed Lemou, Gilles Vilmart.

second-order ODEs, oscillatory solutions, Schrödinger and wave equations, step size restrictions.

In applications to molecular dynamics or quantum dynamics for instance, the right-hand side of (1) involves *fast* forces (short-range interactions) and *slow* forces (long-range interactions). Since *fast* forces are much cheaper to evaluate than *slow* forces, it seems highly desirable to design numerical methods for which the number of evaluations of slow forces is not (at least not too much) affected by the presence of fast forces.

A typical model of highly-oscillatory systems is the second-order differential equations

$$\ddot{q} \quad = \quad -\nabla V(q) \tag{39}$$

where the potential $V(q)$ is a sum of potentials $V = W + U$ acting on different time-scales, with $\nabla^2 W$ positive definite and $\|\nabla^2 W\| >> \|\nabla^2 U\|$. In order to get a bounded error propagation in the linearized equations for an explicit numerical method, the step size must be restricted according to

$$h\omega < C,$$

where C is a constant depending on the numerical method and where ω is the highest frequency of the problem, i.e. in this situation the square root of the largest eigenvalue of $\nabla^2 W$. In applications to molecular dynamics for instance, *fast* forces deriving from W (short-range interactions) are much cheaper to evaluate than *slow* forces deriving from U (long-range interactions). In this case, it thus seems highly desirable to design numerical methods for which the number of evaluations of slow forces is not (at least not too much) affected by the presence of fast forces.

Another prominent example of highly-oscillatory systems is encountered in quantum dynamics where the Schrödinger equation is the model to be used. Assuming that the Laplacian has been discretized in space, one indeed gets the *time*-dependent Schrödinger equation:

$$i\dot{\psi}(t) = \frac{1}{\varepsilon} H(t)\psi(t), \tag{40}$$

where $H(t)$ is finite-dimensional matrix and where ε typically is the square-root of a mass-ratio (say electron/ion for instance) and is small ($\varepsilon \approx 10^{-2}$ or smaller). Through the coupling with classical mechanics ($H(t)$ is obtained by solving some equations from classical mechanics), we are faced once again with two different time-scales, 1 and ε. In this situation also, it is thus desirable to devise a numerical method able to advance the solution by a time-step $h > \varepsilon$.

3.3. Geometric schemes for the Schrödinger equation

Participants: François Castella, Philippe Chartier, Erwan Faou, Florian Méhats, Gilles Vilmart.

Schrödinger equation, variational splitting, energy conservation.

Given the Hamiltonian structure of the Schrödinger equation, we are led to consider the question of energy preservation for time-discretization schemes.

At a higher level, the Schrödinger equation is a partial differential equation which may exhibit Hamiltonian structures. This is the case of the time-dependent Schrödinger equation, which we may write as

$$i\varepsilon\frac{\partial\psi}{\partial t} = H\psi, \tag{41}$$

where $\psi = \psi(x,t)$ is the wave function depending on the spatial variables $x = (x_1, \cdots, x_N)$ with $x_k \in \mathbb{R}^d$ (e.g., with $d = 1$ or 3 in the partition) and the time $t \in \mathbb{R}$. Here, ε is a (small) positive number representing the scaled Planck constant and i is the complex imaginary unit. The Hamiltonian operator H is written

$$H = T + V$$

with the kinetic and potential energy operators

$$T = -\sum_{k=1}^{N}\frac{\varepsilon^2}{2m_k}\Delta_{x_k} \quad \text{and} \quad V = V(x),$$

where $m_k > 0$ is a particle mass and Δ_{x_k} the Laplacian in the variable $x_k \in \mathbb{R}^d$, and where the real-valued potential V acts as a multiplication operator on ψ.

The multiplication by i in (8) plays the role of the multiplication by J in classical mechanics, and the energy $\langle\psi|H|\psi\rangle$ is conserved along the solution of (8), using the physicists' notations $\langle u|A|u\rangle = \langle u, Au\rangle$ where $\langle\ ,\ \rangle$ denotes the Hermitian L^2-product over the phase space. In quantum mechanics, the number N of particles is very large making the direct approximation of (8) very difficult.

The numerical approximation of (8) can be obtained using projections onto submanifolds of the phase space, leading to various PDEs or ODEs: see [64], [63] for reviews. However the long-time behavior of these approximated solutions is well understood only in this latter case, where the dynamics turns out to be finite dimensional. In the general case, it is very difficult to prove the preservation of qualitative properties of (8) such as energy conservation or growth in time of Sobolev norms. The reason for this is that backward error analysis is not directly applicable for PDEs. Overwhelming these difficulties is thus a very interesting challenge.

A particularly interesting case of study is given by symmetric splitting methods, such as the Strang splitting:

$$\psi_1 = \exp\left(-i(\delta t)V/2\right)\exp\left(i(\delta t)\Delta\right)\exp\left(-i(\delta t)V/2\right)\psi_0 \tag{42}$$

where δt is the time increment (we have set all the parameters to 1 in the equation). As the Laplace operator is unbounded, we cannot apply the standard methods used in ODEs to derive long-time properties of these schemes. However, its projection onto finite dimensional submanifolds (such as Gaussian wave packets space or FEM finite dimensional space of functions in x) may exhibit Hamiltonian or Poisson structure, whose long-time properties turn out to be more tractable.

3.4. High-frequency limit of the Helmholtz equation

Participant: François Castella.

waves, Helmholtz equation, high oscillations.

The Helmholtz equation models the propagation of waves in a medium with variable refraction index. It is a simplified version of the Maxwell system for electro-magnetic waves.

The high-frequency regime is characterized by the fact that the typical wavelength of the signals under consideration is much smaller than the typical distance of observation of those signals. Hence, in the high-frequency regime, the Helmholtz equation at once involves highly oscillatory phenomena that are to be described in some asymptotic way. Quantitatively, the Helmholtz equation reads

$$i\alpha_\varepsilon u_\varepsilon(x) + \varepsilon^2 \Delta_x u_\varepsilon + n^2(x) u_\varepsilon = f_\varepsilon(x). \tag{43}$$

Here, ε is the small adimensional parameter that measures the typical wavelength of the signal, $n(x)$ is the space-dependent refraction index, and $f_\varepsilon(x)$ is a given (possibly dependent on ε) source term. The unknown is $u_\varepsilon(x)$. One may think of an antenna emitting waves in the whole space (this is the $f_\varepsilon(x)$), thus creating at any point x the signal $u_\varepsilon(x)$ along the propagation. The small $\alpha_\varepsilon > 0$ term takes into account damping of the waves as they propagate.

One important scientific objective typically is to describe the high-frequency regime in terms of *rays* propagating in the medium, that are possibly refracted at interfaces, or bounce on boundaries, etc. Ultimately, one would like to replace the true numerical resolution of the Helmholtz equation by that of a simpler, asymptotic model, formulated in terms of rays.

In some sense, and in comparison with, say, the wave equation, the specificity of the Helmholtz equation is the following. While the wave equation typically describes the evolution of waves between some initial time and some given observation time, the Helmholtz equation takes into account at once the propagation of waves over *infinitely long* time intervals. Qualitatively, in order to have a good understanding of the signal observed in some bounded region of space, one readily needs to be able to describe the propagative phenomena in the whole space, up to infinity. In other words, the "rays" we refer to above need to be understood from the initial time up to infinity. This is a central difficulty in the analysis of the high-frequency behaviour of the Helmholtz equation.

3.5. From the Schrödinger equation to Boltzmann-like equations

Participant: François Castella.

Schrödinger equation, asymptotic model, Boltzmann equation.

The Schrödinger equation is the appropriate way to describe transport phenomena at the scale of electrons. However, for real devices, it is important to derive models valid at a larger scale.

In semi-conductors, the Schrödinger equation is the ultimate model that allows to obtain quantitative information about electronic transport in crystals. It reads, in convenient adimensional units,

$$i\partial_t \psi(t,x) = -\frac{1}{2}\Delta_x \psi + V(x)\psi, \tag{44}$$

where $V(x)$ is the potential and $\psi(t,x)$ is the time- and space-dependent wave function. However, the size of real devices makes it important to derive simplified models that are valid at a larger scale. Typically, one wishes to have kinetic transport equations. As is well-known, this requirement needs one to be able to describe "collisions" between electrons in these devices, a concept that makes sense at the macroscopic level, while it does not at the microscopic (electronic) level. Quantitatively, the question is the following: can one obtain the Boltzmann equation (an equation that describes collisional phenomena) as an asymptotic model for the Schrödinger equation, along the physically relevant micro-macro asymptotics? From the point of view of modelling, one wishes here to understand what are the "good objects", or, in more technical words, what are the relevant "cross-sections", that describe the elementary collisional phenomena. Quantitatively, the Boltzmann equation reads, in a simplified, linearized, form :

$$\partial_t f(t,x,v) = \int_{\mathbf{R}^3} \sigma(v,v') \, [f(t,x,v') - f(t,x,v)] dv'. \tag{45}$$

Here, the unknown is $f(x,v,t)$, the probability that a particle sits at position x, with a velocity v, at time t. Also, $\sigma(v,v')$ is called the cross-section, and it describes the probability that a particle "jumps" from velocity v to velocity v' (or the converse) after a collision process.

MATHERIALS Team

3. Research Program

3.1. Research Program

Quantum Chemistry aims at understanding the properties of matter through the modeling of its behavior at a subatomic scale, where matter is described as an assembly of nuclei and electrons. At this scale, the equation that rules the interactions between these constitutive elements is the Schrödinger equation. It can be considered (except in few special cases notably those involving relativistic phenomena or nuclear reactions) as a universal model for at least three reasons. First it contains all the physical information of the system under consideration so that any of the properties of this system can in theory be deduced from the Schrödinger equation associated to it. Second, the Schrödinger equation does not involve any empirical parameters, except some fundamental constants of Physics (the Planck constant, the mass and charge of the electron, ...); it can thus be written for any kind of molecular system provided its chemical composition, in terms of natures of nuclei and number of electrons, is known. Third, this model enjoys remarkable predictive capabilities, as confirmed by comparisons with a large amount of experimental data of various types. On the other hand, using this high quality model requires working with space and time scales which are both very tiny: the typical size of the electronic cloud of an isolated atom is the Angström (10^{-10} meters), and the size of the nucleus embedded in it is 10^{-15} meters; the typical vibration period of a molecular bond is the femtosecond (10^{-15} seconds), and the characteristic relaxation time for an electron is 10^{-18} seconds. Consequently, Quantum Chemistry calculations concern very short time (say 10^{-12} seconds) behaviors of very small size (say 10^{-27} m^3) systems. The underlying question is therefore whether information on phenomena at these scales is useful in understanding or, better, predicting macroscopic properties of matter. It is certainly not true that *all* macroscopic properties can be simply upscaled from the consideration of the short time behavior of a tiny sample of matter. Many of them derive from ensemble or bulk effects, that are far from being easy to understand and to model. Striking examples are found in solid state materials or biological systems. Cleavage, the ability of minerals to naturally split along crystal surfaces (e.g. mica yields to thin flakes), is an ensemble effect. Protein folding is also an ensemble effect that originates from the presence of the surrounding medium; it is responsible for peculiar properties (e.g. unexpected acidity of some reactive site enhanced by special interactions) upon which vital processes are based. However, it is undoubtedly true that *many* macroscopic phenomena originate from elementary processes which take place at the atomic scale. Let us mention for instance the fact that the elastic constants of a perfect crystal or the color of a chemical compound (which is related to the wavelengths absorbed or emitted during optic transitions between electronic levels) can be evaluated by atomic scale calculations. In the same fashion, the lubricative properties of graphite are essentially due to a phenomenon which can be entirely modeled at the atomic scale. It is therefore reasonable to simulate the behavior of matter at the atomic scale in order to understand what is going on at the macroscopic one. The journey is however a long one. Starting from the basic principles of Quantum Mechanics to model the matter at the subatomic scale, one finally uses statistical mechanics to reach the macroscopic scale. It is often necessary to rely on intermediate steps to deal with phenomena which take place on various *mesoscales*. It may then be possible to couple one description of the system with some others within the so-called *multiscale* models. The sequel indicates how this journey can be completed focusing on the first smallest scales (the subatomic one), rather than on the larger ones. It has already been mentioned that at the subatomic scale, the behavior of nuclei and electrons is governed by the Schrödinger equation, either in its time dependent form or in its time independent form. Let us only mention at this point that

- both equations involve the quantum Hamiltonian of the molecular system under consideration; from a mathematical viewpoint, it is a self-adjoint operator on some Hilbert space; *both* the Hilbert space and the Hamiltonian operator depend on the nature of the system;
- also present into these equations is the wavefunction of the system; it completely describes its state; its L^2 norm is set to one.

The time dependent equation is a first order linear evolution equation, whereas the time-independent equation is a linear eigenvalue equation. For the reader more familiar with numerical analysis than with quantum mechanics, the linear nature of the problems stated above may look auspicious. What makes the numerical simulation of these equations extremely difficult is essentially the huge size of the Hilbert space: indeed, this space is roughly some symmetry-constrained subspace of $L^2(\mathbb{R}^d)$, with $d = 3(M + N)$, M and N respectively denoting the number of nuclei and the number of electrons the system is made of. The parameter d is already 39 for a single water molecule and rapidly reaches 10^6 for polymers or biological molecules. In addition, a consequence of the universality of the model is that one has to deal at the same time with several energy scales. In molecular systems, the basic elementary interaction between nuclei and electrons (the two-body Coulomb interaction) appears in various complex physical and chemical phenomena whose characteristic energies cover several orders of magnitude: the binding energy of core electrons in heavy atoms is 10^4 times as large as a typical covalent bond energy, which is itself around 20 times as large as the energy of a hydrogen bond. High precision or at least controlled error cancellations are thus required to reach chemical accuracy when starting from the Schrödinger equation. Clever approximations of the Schrödinger problems are therefore needed. The main two approximation strategies, namely the Born-Oppenheimer-Hartree-Fock and the Born-Oppenheimer-Kohn-Sham strategies, end up with large systems of coupled *nonlinear* partial differential equations, each of these equations being posed on $L^2(\mathbb{R}^3)$. The size of the underlying functional space is thus reduced at the cost of a dramatic increase of the mathematical complexity of the problem: nonlinearity. The mathematical and numerical analysis of the resulting models has been the major concern of the team for a long time. In the recent years, while part of the activity still follows this path, the focus has progressively shifted to problems at other scales. Such problems are described in the following sections.

MATHRISK Project-Team

3. Research Program

3.1. Dependence modeling

Participants: Aurélien Alfonsi, Benjamin Jourdain, Damien Lamberton, Bernard Lapeyre.

The volatility is a key concept in modern mathematical finance, and an indicator of the market stability. Risk management and associated instruments depend strongly on the volatility, and volatility modeling has thus become a crucial issue in the finance industry. Of particular importance is the assets *dependence* modeling. The calibration of models for a single asset can now be well managed by banks but modeling of dependence is the bottleneck to efficiently aggregate such models. A typical issue is how to go from the individual evolution of each stock belonging to an index to the joint modeling of these stocks. In this perspective, we want to model stochastic volatility in a *multidimensional* framework. To handle these questions mathematically, we have to deal with stochastic differential equations that are defined on matrices in order to model either the instantaneous covariance or the instantaneous correlation between the assets. From a numerical point of view, such models are very demanding since the main indexes include generally more than thirty assets. It is therefore necessary to develop efficient numerical methods for pricing options and calibrating such models to market data. As a first application, modeling the dependence between assets allows us to better handle derivatives products on a basket. It would give also a way to price and hedge consistenly single-asset and basket products. Besides, it can be a way to capture how the market estimates the dependence between assets. This could give some insights on how the market anticipates the systemic risk.

3.2. Liquidity risk

Participants: Aurélien Alfonsi, Agnès Sulem, Antonino Zanette.

The financial crisis has caused an increased interest in mathematical finance studies which take into account the market incompleteness issue and the liquidity risk. Loosely speaking, liquidity risk is the risk that comes from the difficulty of selling (or buying) an asset. At the extreme, this may be the impossibility to sell an asset, which occurred for "junk assets" during the subprime crisis. Hopefully, it is in general possible to sell assets, but this may have some cost. Let us be more precise. Usually, assets are quoted on a market with a Limit Order Book (LOB) that registers all the waiting limit buy and sell orders for this asset. The bid (resp. ask) price is the most expensive (resp. cheapest) waiting buy or sell order. If a trader wants to sell a single asset, he will sell it at the bid price. Instead, if he wants to sell a large quantity of assets, he will have to sell them at a lower price in order to match further waiting buy orders. This creates an extra cost, and raises important issues. From a short-term perspective (from few minutes to some days), this may be interesting to split the selling order and to focus on finding optimal selling strategies. This requires to model the market microstructure, i.e. how the market reacts in a short time-scale to execution orders. From a long-term perspective (typically, one month or more), one has to understand how this cost modifies portfolio managing strategies (especially delta-hedging or optimal investment strategies). At this time-scale, there is no need to model precisely the market microstructure, but one has to specify how the liquidity costs aggregate.

3.2.1. Long term liquidity risk.

On a long-term perspective, illiquidity can be approached via various ways: transactions costs [41], [42], [53], [61], [67], [87], [83], delay in the execution of the trading orders [88], [86], [57], trading constraints or restriction on the observation times (see e.g. [63] and references herein). As far as derivative products are concerned, one has to understand how delta-hedging strategies have to be modified. This has been considered for example by Cetin, Jarrow and Protter [85]. We plan to contribute on these various aspects of liquidity risk modeling and associated stochastic optimization problems. Let us mention here that the price impact generated by the trades of the investor is often neglected with a long-term perspective. This seems acceptable

since the investor has time enough to trade slowly in order to eliminate its market impact. Instead, when the investor wants to make significant trades on a very short time horizon, it is crucial to take into account and to model how prices are modified by these trades. This question is addressed in the next paragraph on market microstructure.

3.2.2. Market microstructure.

The European directive MIFID has increased the competition between markets (NYSE-Euronext, Nasdaq, LSE and new competitors). As a consequence, the cost of posting buy or sell orders on markets has decreased, which has stimulated the growth of market makers. Market makers are posting simultaneously bid and ask orders on a same stock, and their profit comes from the bid-ask spread. Basically, their strategy is a "round-trip" (i.e. their position is unchanged between the beginning and the end of the day) that has generated a positive cash flow.

These new rules have also greatly stimulated research on market microstructure modeling. From a practitioner point of view, the main issue is to solve the so-called "optimal execution problem": given a deadline T, what is the optimal strategy to buy (or sell) a given amount of shares that achieves the minimal expected cost? For large amounts, it may be optimal to split the order into smaller ones. This is of course a crucial issue for brokers, but also market makers that are looking for the optimal round-trip.

Solving the optimal execution problem is not only an interesting mathematical challenge. It is also a mean to better understand market viability, high frequency arbitrage strategies and consequences of the competition between markets. For example when modeling the market microstructure, one would like to find conditions that allow or exclude round trips. Beyond this, even if round trips are excluded, it can happen that an optimal selling strategy is made with large intermediate buy trades, which is unlikely and may lead to market instability.

We are interested in finding synthetic market models in which we can describe and solve the optimal execution problem. A. Alfonsi and A. Schied (Mannheim University) [45] have already proposed a simple Limit Order Book model (LOB) in which an explicit solution can be found for the optimal execution problem. We are now interested in considering more sophisticated models that take into account realistic features of the market such as short memory or stochastic LOB. This is mid term objective. At a long term perspective one would like to bridge these models to the different agent behaviors, in order to understand the effect of the different quotation mechanisms (transaction costs for limit orders, tick size, etc.) on the market stability.

3.3. Contagion modeling and systemic risk

Participants: Benjamin Jourdain, Agnès Sulem.

After the recent financial crisis, systemic risk has emerged as one of the major research topics in mathematical finance. The scope is to understand and model how the bankruptcy of a bank (or a large company) may or not induce other bankruptcies. By contrast with the traditional approach in risk management, the focus is no longer on modeling the risks faced by a single financial institution, but on modeling the complex interrelations between financial institutions and the mechanisms of distress propagation among these. Ideally, one would like to be able to find capital requirements (such as the one proposed by the Basel committee) that ensure that the probability of multiple defaults is below some level.

The mathematical modeling of default contagion, by which an economic shock causing initial losses and default of a few institutions is amplified due to complex linkages, leading to large scale defaults, can be addressed by various techniques, such as network approaches (see in particular R. Cont et al. [46] and A. Minca [72]) or mean field interaction models (Garnier-Papanicolaou-Yang [62]). The recent approach in [46] seems very promising. It describes the financial network approach as a weighted directed graph, in which nodes represent financial institutions and edges the exposures between them. Distress propagation in a financial system may be modeled as an epidemics on this graph. In the case of incomplete information on the structure of the interbank network, cascade dynamics may be reduced to the evolution of a multi-dimensional Markov chain that corresponds to a sequential discovery of exposures and determines at any time the size of contagion. Little has been done so far on the *control* of such systems in order to reduce the systemic risk and we aim to contribute to this domain.

3.4. Stochastic analysis and numerical probability

3.4.1. Stochastic control

Participants: Vlad Bally, Jean-Philippe Chancelier, Marie-Claire Quenez, Agnès Sulem.

The financial crisis has caused an increased interest in mathematical finance studies which take into account the market incompleteness issue and the default risk modeling, the interplay between information and performance, the model uncertainty and the associated robustness questions, and various nonlinearities. We address these questions by further developing the theory of stochastic control in a broad sense, including stochastic optimization, nonlinear expectations, Malliavin calculus, stochastic differential games and various aspects of optimal stopping.

3.4.2. Optimal stopping

Participants: Aurélien Alfonsi, Benjamin Jourdain, Damien Lamberton, Agnès Sulem, Marie-Claire Quenez.

The theory of American option pricing has been an incite for a number of research articles about optimal stopping. Our recent contributions in this field concern optimal stopping in models with jumps, irregular obstacles, free boundary analysis, reflected BSDEs.

3.4.3. Simulation of stochastic differential equations

Participants: Benjamin Jourdain, Aurélien Alfonsi, Vlad Bally, Damien Lamberton, Bernard Lapeyre, Jérôme Lelong, Céline Labart.

Effective numerical methods are crucial in the pricing and hedging of derivative securities. The need for more complex models leads to stochastic differential equations which cannot be solved explicitly, and the development of discretization techniques is essential in the treatment of these models. The project MathRisk addresses fundamental mathematical questions as well as numerical issues in the following (non exhaustive) list of topics: Multidimensional stochastic differential equations, High order discretization schemes, Singular stochastic differential equations, Backward stochastic differential equations.

3.4.4. Monte-Carlo simulations

Participants: Benjamin Jourdain, Aurélien Alfonsi, Damien Lamberton, Vlad Bally, Bernard Lapeyre, Ahmed Kebaier, Céline Labart, Jérôme Lelong, Antonino Zanette.

Monte-Carlo methods is a very useful tool to evaluate prices especially for complex models or options. We carry on research on *adaptive variance reduction methods* and to use *Monte-Carlo methods for calibration* of advanced models.

This activity in the MathRisk team is strongly related to the development of the Premia software.

3.4.5. Malliavin calculus and applications in finance

Participants: Vlad Bally, Arturo Kohatsu-Higa, Agnès Sulem, Antonino Zanette.

The original Stochastic Calculus of Variations, now called the Malliavin calculus, was developed by Paul Malliavin in 1976 [70]. It was originally designed to study the smoothness of the densities of solutions of stochastic differential equations. One of its striking features is that it provides a probabilistic proof of the celebrated Hörmander theorem, which gives a condition for a partial differential operator to be hypoelliptic. This illustrates the power of this calculus. In the following years a lot of probabilists worked on this topic and the theory was developed further either as analysis on the Wiener space or in a white noise setting. Many applications in the field of stochastic calculus followed. Several monographs and lecture notes (for example D. Nualart [74], D. Bell [52] D. Ocone [76], B. Øksendal [89]) give expositions of the subject. See also V. Bally [47] for an introduction to Malliavin calculus.

From the beginning of the nineties, applications of the Malliavin calculus in finance have appeared : In 1991 Karatzas and Ocone showed how the Malliavin calculus, as further developed by Ocone and others, could be used in the computation of hedging portfolios in complete markets [75].

Since then, the Malliavin calculus has raised increasing interest and subsequently many other applications to finance have been found [71], such as minimal variance hedging and Monte Carlo methods for option pricing. More recently, the Malliavin calculus has also become a useful tool for studying insider trading models and some extended market models driven by Lévy processes or fractional Brownian motion.

We give below an idea why Malliavin calculus may be a useful instrument for probabilistic numerical methods.

We recall that the theory is based on an integration by parts formula of the form $E(f'(X)) = E(f(X)Q)$. Here X is a random variable which is supposed to be "smooth" in a certain sense and non-degenerated. A basic example is to take $X = \sigma\Delta$ where Δ is a standard normally distributed random variable and σ is a strictly positive number. Note that an integration by parts formula may be obtained just by using the usual integration by parts in the presence of the Gaussian density. But we may go further and take X to be an aggregate of Gaussian random variables (think for example of the Euler scheme for a diffusion process) or the limit of such simple functionals.

An important feature is that one has a relatively explicit expression for the weight Q which appears in the integration by parts formula, and this expression is given in terms of some Malliavin-derivative operators.

Let us now look at one of the main consequences of the integration by parts formula. If one considers the *Dirac* function $\delta_x(y)$, then $\delta_x(y) = H'(y - x)$ where H is the *Heaviside* function and the above integration by parts formula reads $E(\delta_x(X)) = E(H(X - x)Q)$, where $E(\delta_x(X))$ can be interpreted as the density of the random variable X. We thus obtain an integral representation of the density of the law of X. This is the starting point of the approach to the density of the law of a diffusion process: the above integral representation allows us to prove that under appropriate hypothesis the density of X is smooth and also to derive upper and lower bounds for it. Concerning simulation by Monte Carlo methods, suppose that you want to compute $E(\delta_x(y)) \sim \frac{1}{M}\sum_{i=1}^{M} \delta_x(X^i)$ where $X^1, ..., X^M$ is a sample of X. As X has a law which is absolutely continuous with respect to the Lebesgue measure, this will fail because no X^i hits exactly x. But if you are able to simulate the weight Q as well (and this is the case in many applications because of the explicit form mentioned above) then you may try to compute $E(\delta_x(X)) = E(H(X - x)Q) \sim \frac{1}{M}\sum_{i=1}^{M} E(H(X^i - x)Q^i)$. This basic remark formula leads to efficient methods to compute by a Monte Carlo method some irregular quantities as derivatives of option prices with respect to some parameters (the *Greeks*) or conditional expectations, which appear in the pricing of American options by the dynamic programming). See the papers by Fournié et al [60] and [59] and the papers by Bally et al., Benhamou, Bermin et al., Bernis et al., Cvitanic et al., Talay and Zheng and Temam in [69].

L. Caramellino, A. Zanette and V. Bally have been concerned with the computation of conditional expectations using Integration by Parts formulas and applications to the numerical computation of the price and the Greeks (sensitivities) of American or Bermudean options. The aim of this research was to extend a paper of Reigner and Lions who treated the problem in dimension one to higher dimension - which represent the real challenge in this field. Significant results have been obtained up to dimension 5 [51] and the corresponding algorithms have been implemented in the Premia software.

Moreover, there is an increasing interest in considering jump components in the financial models, especially motivated by calibration reasons. Algorithms based on the integration by parts formulas have been developed in order to compute Greeks for options with discontinuous payoff (e.g. digital options). Several papers and two theses (M. Messaoud and M. Bavouzet defended in 2006) have been published on this topic and the corresponding algorithms have been implemented in Premia. Malliavin Calculus for jump type diffusions - and more general for random variables with locally smooth law - represents a large field of research, also for applications to credit risk problems.

The Malliavin calculus is also used in models of insider trading. The "enlargement of filtration" technique plays an important role in the modeling of such problems and the Malliavin calculus can be used to obtain general results about when and how such filtration enlargement is possible. See the paper by P. Imkeller in [69]). Moreover, in the case when the additional information of the insider is generated by adding the information about the value of one extra random variable, the Malliavin calculus can be used to find explicitly the optimal

portfolio of an insider for a utility optimization problem with logarithmic utility. See the paper by J.A. León, R. Navarro and D. Nualart in [69]).

A. Kohatsu Higa and A. Sulem have studied a controlled stochastic system whose state is described by a stochastic differential equation with anticipating coefficients. These SDEs can be interpreted in the sense of *forward integrals*, which are the natural generalization of the semimartingale integrals, as introduced by Russo and Valois [82]. This methodology has been applied for utility maximization with insiders.

Maxplus Project-Team

3. Research Program

3.1. L'algèbre max-plus/Max-plus algebra

Le semi-corps *max-plus* est l'ensemble $\mathbb{R} \cup \{-\infty\}$, muni de l'addition $(a, b) \mapsto a \oplus b = \max(a, b)$ et de la multiplication $(a, b) \mapsto a \otimes b = a + b$. Cette structure algébrique diffère des structures de corps classiques par le fait que l'addition n'est pas une loi de groupe, mais est idempotente: $a \oplus a = a$. On rencontre parfois des variantes de cette structure: par exemple, le semi-corps *min-plus* est l'ensemble $\mathbb{R} \cup \{+\infty\}$ muni des lois $a \oplus b = \min(a, b)$ et $a \otimes b = a + b$, et le semi-anneau *tropical* est l'ensemble $\mathbb{N} \cup \{+\infty\}$ munis des mêmes lois. L'on peut se poser la question de généraliser les constructions de l'algèbre et de l'analyse classique, qui reposent pour une bonne part sur des anneaux ou des corps tels que $\mathbb{Z}$ ou $\mathbb{R}$, au cas de semi-anneaux de type max-plus: tel est l'objet de ce qu'on appelle un peu familièrement "l'algèbre max-plus".

Il est impossible ici de donner une vue complète du domaine. Nous nous bornerons à indiquer quelques références bibliographiques. L'intérêt pour les structures de type max-plus est contemporain de la naissance de la théorie des treillis [114]. Depuis, les structures de type max-plus ont été développées indépendamment par plusieurs écoles, en relation avec plusieurs domaines. Les motivations venant de la Recherche Opérationnelle (programmation dynamique, problèmes de plus court chemin, problèmes d'ordonnancement, optimisation discrète) ont été centrales dans le développement du domaine [103], [134], [183], [187], [188]. Les semi-anneaux de type max-plus sont bien sûr reliés aux algèbres de Boole [90]. L'algèbre max-plus apparaît de manière naturelle en contrôle optimal et dans la théorie des équations aux dérivées partielles d'Hamilton-Jacobi [173], [171], [157], [141], [130], [176], [150], [131], [117], [65]. Elle apparaît aussi en analyse asymptotique (asymptotiques de type WKB [156], [157], [141], grandes déviations [170], asymptotiques à température nulle en physique statistique [92]), puisque l'algèbre max-plus apparaît comme limite de l'algèbre usuelle. La théorie des opérateurs linéaires max-plus peut être vue comme faisant partie de la théorie des opérateurs de Perron-Frobenius non-linéaires, ou de la théorie des applications contractantes ou monotones sur les cônes [142], [161], [154], [79], laquelle a de nombreuse motivations, telles l'économie mathématique [159], et la théorie des jeux [174], [54]. Dans la communauté des systèmes à événements discrets, l'algèbre max-plus a été beaucoup étudiée parce qu'elle permet de représenter de manière linéaire les phénomènes de synchronisation, lesquels déterminent le comportement temporel de systèmes de production ou de réseaux, voir [6]. Parmi les développements récents du domaine, on peut citer le calcul des réseaux [91], [146], qui permet de calculer des bornes pire des cas de certaines mesures de qualité de service. En informatique théorique, l'algèbre max-plus (ou plutôt le semi-anneau tropical) a joué un rôle décisif dans la résolution de problèmes de décision en théorie des automates [178], [137], [179], [143], [163]. Notons finalement, pour information, que l'algèbre max-plus est apparue récemment en géométrie algébrique [129], [182], [158], [181] et en théorie des représentations [118], [82], sous les noms de géométrie et combinatoire tropicales.

Nous décrivons maintenant de manière plus détaillée les sujets qui relèvent directement des intérêts du projet, comme la commande optimale, les asymptotiques, et les systèmes à événements discrets.

English version

The *max-plus* semifield is the set $\mathbb{R} \cup \{-\infty\}$, equipped with the addition $(a, b) \mapsto a \oplus b = \max(a, b)$ and the multiplication $(a, b) \mapsto a \otimes b = a + b$. This algebraic structure differs from classical structures, like fields, in that addition is idempotent: $a \oplus a = a$. Several variants have appeared in the literature: for instance, the *min-plus* semifield is the set $\mathbb{R} \cup \{+\infty\}$ equipped with the laws $a \oplus b = \min(a, b)$ and $a \otimes b = a + b$, and the *tropical* semiring is the set $\mathbb{N} \cup \{+\infty\}$ equipped with the same laws. One can ask the question of extending to max-plus type structures the classical constructions and results of algebra and analysis: this is what is often called in a wide sense "max-plus algebra" or "tropical algebra".

It is impossible to give in this short space a fair view of the field. Let us, however, give a few references. The interest in max-plus type structures is contemporaneous with the early developments of lattice theory [114]. Since that time, max-plus structures have been developed independently by several schools, in relation with several fields. Motivations from Operations Research (dynamic programming, shortest path problems, scheduling problems, discrete optimisation) were central in the development of the field [103], [134], [183], [187], [188]. Of course, max-plus type semirings are related to Boolean algebras [90]. Max-plus algebras arises naturally in optimal control and in the theory of Hamilton-Jacobi partial differential equations [173], [171], [157], [141], [130], [176], [150], [131], [117], [65]. It arises in asymptotic analysis (WKB asymptotics [156], [157], [141], large deviation asymptotics [170], or zero temperature asymptotics in statistical physics [92]), since max-plus algebra appears as a limit of the usual algebra. The theory of max-plus linear operators may be thought of as a part of the non-linear Perron-Frobenius theory, or of the theory of nonexpansive or monotone operators on cones [142], [161], [154], [79], a theory with numerous motivations, including mathematical economy [159] and game theory [174], [54]. In the discrete event systems community, max-plus algebra has been much studied since it allows one to represent linearly the synchronisation phenomena which determine the time behaviour of manufacturing systems and networks, see [6]. Recent developments include the network calculus of [91], [146] which allows one to compute worst case bounds for certain measures of quality of service. In theoretical computer science, max-plus algebra (or rather, the tropical semiring) played a key role in the solution of decision problems in automata theory [178], [137], [179], [143], [163]. We finally note for information that max-plus algebra has recently arisen in algebraic geometry [129], [182], [158], [181] and in representation theory [118], [82], under the names of tropical geometry and combinatorics.

We now describe in more details some parts of the subject directly related to our interests, like optimal control, asymptotics, and discrete event systems.

3.2. Algèbre max-plus, programmation dynamique, et commande optimale/Max-plus algebra, dynamic programming, and optimal control

L'exemple le plus simple d'un problème conduisant à une équation min-plus linéaire est le problème classique du plus court chemin. Considérons un graphe dont les nœuds sont numérotés de 1 à n et dont le coût de l'arc allant du nœud i au nœud j est noté $M_{ij} \in \mathbb{R} \cup \{+\infty\}$. Le coût minimal d'un chemin de longueur k, allant de i à j, est donné par la quantité:

$$v_{ij}(k) = \min_{\ell:\ \ell_0=i,\ \ell_k=j} \sum_{r=0}^{k-1} M_{\ell_r \ell_{r+1}} \ , \tag{46}$$

où le minimum est pris sur tous les chemins $\ell = (\ell_0, ..., \ell_k)$ de longueur k, de nœud initial $\ell_0 = i$ et de nœud final $\ell_k = j$. L'équation classique de la programmation dynamique s'écrit:

$$v_{ij}(k) = \min_{1 \le s \le n} (M_{is} + v_{sj}(k-1)) \ . \tag{47}$$

On reconnaît ainsi une équation linéaire min-plus :

$$v(k) = Mv(k-1) \ , \tag{48}$$

où on note par la concaténation le produit matriciel induit par la structure de l'algèbre min-plus. Le classique *problème de Lagrange* du calcul des variations,

$$v(x,T) = \inf_{X(\cdot),\ X(0)=x} \int_0^T L(X(t), \dot{X}(t))\mathrm{d}t + \phi(X(T)) \ , \tag{49}$$

où $X(t) \in \mathbb{R}^n$, pour $0 \leq t \leq T$, et $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ est le Lagrangien, peut être vu comme une version continue de (1), ce qui permet de voir l'équation d'Hamilton-Jacobi que vérifie v,

$$v(\cdot, 0) = \phi, \quad \frac{\partial v}{\partial T} + H(x, \frac{\partial v}{\partial x}) = 0, \qquad H(x,p) = \sup_{y \in \mathbb{R}^n} (-p \cdot y - L(x,y)) \ , \tag{50}$$

comme une équation min-plus linéaire. En particulier, les solutions de (5) vérifient un principe de superposition min-plus: si v et w sont deux solutions, et si $\lambda, \mu \in \mathbb{R}$, $\inf(\lambda + v, \mu + w)$ est encore solution de (5). Ce point de vue, inauguré par Maslov, a conduit au développement de l'école d'Analyse Idempotente (voir [157], [141], [150]).

La présence d'une structure algébrique sous-jacente permet de voir les solutions stationnaires de (2) et (5) comme des vecteurs propres de la matrice M ou du semi-groupe d'évolution de l'équation d'Hamilton-Jacobi. La valeur propre associée fournit le coût moyen par unité de temps (coût ergodique). La représentation des vecteurs propres (voir [173], [183], [103], [132], [97], [78], [6] pour la dimension finie, et [157], [141] pour la dimension infinie) est intimement liée au théorème de l'autoroute qui décrit les trajectoires optimales quand la durée ou la longueur des chemins tend vers l'infini. Pour l'équation d'Hamilton-Jacobi, des résultats reliés sont apparus récemment en théorie d'"Aubry-Mather" [117].

English version

The most elementary example of a problem leading to a min-plus linear equation is the classical shortest path problem. Consider a graph with nodes $1, ..., n$, and let $M_{ij} \in \mathbb{R} \cup \{+\infty\}$ denote the cost of the arc from node i to node j. The minimal cost of a path of a given length, k, from i to j, is given by (1), where the minimum is taken over all paths $\ell = (\ell_0, ..., \ell_k)$ of length k, with initial node $\ell_0 = i$ and final node $\ell_k = j$. The classical dynamic programming equation can be written as in (2). We recognise the min-plus linear equation (3), where concatenation denotes the matrix product induced by the min-plus algebraic structure. The classical *Lagrange problem* of calculus of variations, given by (4) where $X(t) \in \mathbb{R}^n$, for $0 \leq t \leq T$, and $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the Lagrangian, may be thought of as a continuous version of (1), which allows us to see the Hamilton-Jacobi equation (5) satisfied by v, as a min-plus linear equation. In particular, the solutions of (5) satisfy a min-plus superposition principle: if v and w are two solutions, and if $\lambda, \mu \in \mathbb{R}$, then $\inf(\lambda + v, \mu + w)$ is also a solution of (5). This point of view, due to Maslov, led to the developpement of the school of Idempotent Analysis (see [157], [141], [150]).

The underlying algebraic structure allows one to see stationnary solutions of (2) and (5) as eigenvectors of the matrix M or of the evolution semigroup of the Hamilton-Jacobi equation. The associated eigenvalue gives the average cost per time unit (ergodic cost). The representation of eigenvectors (see [173], [183], [132], [97], [103], [78], [6] for the finite dimension case, and [157], [141] for the infinite dimension case) is intimately related to turnpike theorems, which describe optimal trajectories as the horizon, or path length, tends to infinity. For the Hamilton-Jacobi equation, related results have appeared recently in the "Aubry-Mather" theory [117].

3.3. Applications monotones et théorie de Perron-Frobenius non-linéaire, ou l'approche opératorielle du contrôle optimal et des jeux/Monotone maps and non-linear Perron-Frobenius theory, or the operator approach to optimal control and games

On sait depuis le tout début des travaux en décision markovienne que les opérateurs de la programmation dynamique f de problèmes de contrôle optimal ou de jeux (à somme nulle et deux joueurs), avec critère additif, ont les propriétés suivantes :

$$\begin{array}{lrcl} \text{monotonie}/\textit{monotonicity} & x \leq y & \Rightarrow & f(x) \leq f(y) \ , \\ \text{contraction}/\textit{nonexpansiveness} & \|f(x) - f(y)\|_\infty & \leq & \|x - y\|_\infty \ . \end{array} \tag{51}$$

Ici, l'opérateur f est une application d'un certain espace de fonctions à valeurs réelles dans lui-même, $\leq$ désigne l'ordre partiel usuel, et $\|\cdot\|_\infty$ désigne la norme sup. Dans le cas le plus simple, l'ensemble des états est $\{1,...,n\}$ et f est une application de $\mathbb{R}^n$ dans lui-même. Les applications monotones qui sont contractantes pour la norme du sup peuvent être vues comme des généralisations non-linéaires des matrices sous-stochastiques. Une sous-classe utile, généralisant les matrices stochastiques, est formée des applications qui sont monotones et commutent avec l'addition d'une constante [102] (celles ci sont parfois appelées fonctions topicales). Les problèmes de programmation dynamique peuvent être traduits en termes d'opérateurs : l'équation de la programmation dynamique d'un problème de commande optimale à horizon fini s'écrit en effet $x(k) = f(x(k-1))$, où $x(k)$ est la fonction valeur en horizon k et $x(0)$ est donné; la fonction valeur y d'un problème à horizon infini (y compris le cas d'un problème d'arrêt optimal) vérifie $y = f(y)$; la fonction valeur z d'un problème avec facteur d'actualisation $0 < \alpha < 1$ vérifie $z = f(\alpha z)$, etc. Ce point de vue abstrait a été très fructueux, voir par exemple [54]. Il permet d'inclure la programmation dynamique dans la perspective plus large de la théorie de Perron-Frobenius non-linéaire, qui, depuis l'extension du théorème de Perron-Frobenius par Krein et Rutman, traite des applications non linéaires sur des cônes vérifiant des conditions de monotonie, de contraction ou d'homogénéité. Les problèmes auxquels on s'intéresse typiquement sont la structure de l'ensemble des points fixes de f, le comportement asymptotique de f^k, en particulier l'existence de la limite de $f^k(x)/k$ lorsque k tends vers l'infini (afin d'obtenir le coût ergodique d'un problème de contrôle optimal ou de jeux), l'asymptotique plus précise de f^k, à une normalisation près (afin d'obtenir le comportement précis de l'itération sur les valeurs), etc. Nous renvoyons le lecteur à [161] pour un panorama. Signalons que dans [123],[7], des algorithmes inspirés de l'algorithme classique d'itérations sur les politiques du contrôle stochastique ont pu être introduits dans le cas des opérateurs monotones contractants généraux, en utilisant des résultats de structure de l'ensemble des points fixes de ces opérateurs. Les applications de la théorie des applications monotones contractantes ne se limitent pas au contrôle optimal et aux jeux. En particulier, on utilise la même classe d'applications dans la modélisation des systèmes à événements discrets, voir le §3.5 ci-dessous, et une classe semblable d'applications en analyse statique de programmes, voir §4.4 ci-dessous.

English version

Since the very beginning of Markov decision theory, it has been observed that dynamic programming operators f arising in optimal control or (zero-sum, two player) game problems have Properties (6). Here, the operator f is a self-map of a certain space of real valued functions, equipped with the standard ordering $\leq$ and with the sup-norm $\|\cdot\|_\infty$. In the simplest case, the set of states is $\{1,...,n\}$, and f is a self-map of $\mathbb{R}^n$. Monotone maps that are nonexpansive in the sup norm may be thought of as nonlinear generalisations of substochastic matrices. A useful subclass, which generalises stochastic matrices, consists of those maps which are monotone and commute with the addition of a constant [102] (these maps are sometimes called topical functions). Dynamic programming problems can be translated in operator terms: the dynamic programming equation for a finite horizon problem can be written as $x(k) = f(x(k-1))$, where $x(k)$ is the value function in horizon k and $x(0)$ is given; the value function y of a problem with an infinite horizon (including the case of optimal stopping) satisfies $y = f(y)$; the value function z of a problem with discount factor $0 < \alpha < 1$ satisfies $z = f(\alpha z)$, etc. This abstract point of view has been very fruitful, see for instance [54]. It allows one to put dynamic programming in the wider perspective of nonlinear Perron-Frobenius theory, which, after the extension of the Perron-Frobenius theorem by Krein and Rutman, studies non-linear self-maps of cones, satisfying various monotonicity, nonexpansiveness, and homogeneity conditions. Typical problems of interests are the structure of the fixed point set of f, the asymptotic behaviour of f^k, including the existence of the limit of $f^k(x)/k$ as k tends to infinity (which yields the ergodic cost in control or games problems), the finer asymptotic behaviour of f^k, possibly up to a normalisation (which yields precise results on value iteration), etc. We shall not attempt to survey this theory here, and will only refer the reader to [161] for more background. In [123],[7], algorithms inspired from the classical policy iterations algorithm of stochastic control have been introduced for general monotone nonexpansive operators, using structural results for the fixed point set of these operators. Applications of monotone or nonexpansive maps are not limited to optimal control and game theory. In particular, we also use the same class of maps as models of discrete event dynamics systems,

see §3.5 below, and we shall see in §4.4 that related classes of maps are useful in the static analysis of computer programs.

3.4. Processus de Bellman/Bellman processes

Un autre point de vue sur la commande optimale est la théorie des *processus de Bellman* [171], [107], [106], [65],[1], qui fournit un analogue max-plus de la théorie des probabilités. Cette théorie a été développée à partir de la notion de *mesure idempotente* introduite par Maslov [156]. Elle établit une correspondance entre probabilités et optimisation, dans laquelle les variables aléatoires deviennent des variables de coût (qui permettent de paramétriser les problèmes d'optimisation), la notion d'espérance conditionnelle est remplacée par celle de coût conditionnel (pris sur un ensemble de solutions faisables), la propriété de Markov correspond au principe de la programmation dynamique de Bellman, et la convergence faible à une convergence de type épigraphe. Les théorèmes limites pour les processus de Bellman (loi des grands nombres, théorème de la limite centrale, lois stables) fournissent des résultats asymptotiques en commande optimale. Ces résultats généraux permettent en particulier de comprendre qualitativement les difficultés d'approximation des solutions d'équations d'Hamilton-Jacobi retrouvés en particulier dans le travail de thèse d'Asma Lakhoua [144], [62].

English version

Another point of view on optimal control is the theory of *Bellman processes* [171], [107], [106], [65], [1] which provides a max-plus analogue of probability theory, relying on the theory of *idempotent measures* due to Maslov [156]. This establishes a correspondence between probability and optimisation, in which random variables become cost variables (which allow to parametrise optimisation problems), the notion of conditional expectation is replaced by a notion of conditional cost (taken over a subset of feasible solutions), the Markov property corresponds to the Bellman's dynamic programming principle, and weak convergence corresponds to an epigraph-type convergence. Limit theorems for Bellman processes (law of large numbers, central limit theorems, stable laws) yield asymptotic results in optimal control. Such general results help in particular to understand qualitatively the difficulty of approximation of Hamilton-Jacobi equations found again in particular in the PhD thesis work of Asma Lakhoua [144], [62].

3.5. Systèmes à événements discrets/Discrete event systems

Des systèmes dynamiques max-plus linéaires, de type (2), interviennent aussi, avec une interprétation toute différente, dans la modélisation des systèmes à événements discrets. Dans ce contexte, on associe à chaque tâche répétitive, i, une fonction *compteur*, $v_i : \mathbb{R} \to \mathbb{N}$, telle que $v_i(t)$ compte le nombre cumulé d'occurrences de la tâche i jusqu'à l'instant t. Par exemple, dans un système de production, $v_i(t)$ compte le nombre de pièces d'un certain type produites jusqu'à l'instant t. Dans le cas le plus simple, qui dans le langage des réseaux de Petri, correspond à la sous-classe très étudiée des graphes d'événements temporisés [93], on obtient des équations min-plus linéaires analogues à (2). Cette observation, ou plutôt, l'observation duale faisant intervenir des fonctions dateurs, a été le point de départ [97] de l'approche max-plus des systèmes à événements discrets [6], qui fournit un analogue max-plus de la théorie des systèmes linéaires classiques, incluant les notions de représentation d'état, de stabilité, de séries de transfert, etc. En particulier, les valeurs propres fournissent des mesures de performance telles que le taux de production. Des généralisations non-linéaires, telles que les systèmes dynamiques min-max [162], [136], ont aussi été étudiées. Les systèmes dynamiques max-plus linéaires aléatoires sont particulièrement utiles dans la modélisation des réseaux [77]. Les modèles d'automates à multiplicités max-plus [121], incluant certains versions temporisées des modèles de traces ou de tas de pièces [125], permettent de représenter des phénomènes de concurrence ou de partage de ressources. Les automates à multiplicités max-plus on été très étudiés par ailleurs en informatique théorique [178], [137], [149], [179], [143], [163]. Ils fournissent des modèles particulièrement adaptés à l'analyse de problèmes d'ordonnancement [148].

English version

Dynamical systems of type (2) also arise, with a different interpretation, in the modelling of discrete event systems. In this context, one associates to every repetitive task, i, a counter function, $v_i : \mathbb{R} \to \mathbb{N}$, such that $v_i(t)$ gives the total number of occurrences of task i up to time t. For instance, in a manufacturing system, $v_i(t)$ will count the number of parts of a given type produced up to time t. In the simplest case, which, in the vocabulary of Petri nets, corresponds to the much studied subclass of timed event graphs [93], we get min-plus linear equations similar to (2). This observation, or rather, the dual observation concerning dater functions, was the starting point [97] of the max-plus approach of discrete event systems [6], which provides some analogue of the classical linear control theory, including notions of state space representations, stability, transfer series, etc. In particular, eigenvalues yield performance measures like the throughput. Nonlinear generalisations, like min-max dynamical systems [162], [136], have been particularly studied. Random max-plus linear dynamical systems are particularly useful in the modelling of networks [77]. Max-plus automata models [121], which include some timed version of trace or heaps of pieces models [125], allow to represent phenomena of concurrency or resource sharing. Note that max-plus automata have been much studied in theoretical computer science [178], [137], [149], [179], [143], [163]. Such automata models are particularly adapted to the analysis of scheduling problems [148].

3.6. Algèbre linéaire max-plus/Basic max-plus algebra

Une bonne partie des résultats de l'algèbre max-plus concerne l'étude des systèmes d'équations linéaires. On peut distinguer trois familles d'équations, qui sont traitées par des techniques différentes : 1) Nous avons déjà évoqué dans les sections 3.2 et 3.3 le problème spectral max-plus $Ax = \lambda x$ et ses généralisations. Celui-ci apparaît en contrôle optimal déterministe et dans l'analyse des systèmes à événements discrets. 2) Le problème $Ax = b$ intervient en commande juste-à-temps (dans ce contexte, le vecteur x représente les dates de démarrage des tâches initiales, b représente certaines dates limites, et on se contente souvent de l'inégalité $Ax \leq b$). Le problème $Ax = b$ est intimement lié au problème d'affectation optimale, et plus généralement au problème de transport optimal. Il se traite via la théorie des correspondances de Galois abstraites, ou théorie de la résiduation [114], [84], [183], [187],[6]. Les versions dimension infinie du problème $Ax = b$ sont reliées aux questions d'analyse convexe abstraite [180], [175], [60] et de dualité non convexe. 3) Le problème linéaire général $Ax = Bx$ conduit à des développements combinatoires intéressants (polyèdres max-plus, déterminants max-plus, symétrisation [135], [164],[6]). Le sujet fait l'objet d'un intérêt récemment renouvelé [108].

English version

An important class of results in max-plus algebra concerns the study of max-plus linear equations. One can distinguish three families of equations, which are handled using different techniques: 1) We already mentioned in Sections 3.2 and 3.3 the max-plus spectral problem $Ax = \lambda x$ and its generalisations, which appears in deterministic optimal control and in performance analysis of discrete event systems. 2) The $Ax = b$ problem arises naturally in just in time problems (in this context, the vector x represents the starting times of initial tasks, b represents some deadlines, and one is often content with the inequality $Ax \leq b$). The $Ax = b$ problem is intimately related with optimal assignment, and more generally, with optimal transportation problems. Its theory relies on abstract Galois correspondences, or residuation theory [114], [84], [183], [187],[6]. Infinite dimensional versions of the $Ax = b$ problem are related to questions of abstract convex analysis [180], [175], [60] and nonconvex duality. 3) The general linear system $Ax = Bx$ leads to interesting combinatorial developments (max-plus polyhedra, determinants, symmetrisation [135], [164],[6]). The subject has attracted recently a new attention [108].

3.7. Algèbre max-plus et asymptotiques/Using max-plus algebra in asymptotic analysis

Le rôle de l'algèbre min-plus ou max-plus dans les problèmes asymptotiques est évident si l'on écrit

$$e^{-a/\epsilon} + e^{-b/\epsilon} \asymp e^{-\min(a,b)/\epsilon} \ , \qquad e^{-a/\epsilon} \times e^{-b/\epsilon} = e^{-(a+b)/\epsilon} \ , \tag{52}$$

lorsque $\epsilon \to 0^+$. Formellement, l'algèbre min-plus peut être vue comme la limite d'une déformation de l'algèbre classique, en introduisant le semi-anneau $\mathbb{R}_\epsilon$, qui est l'ensemble $\mathbb{R} \cup \{+\infty\}$, muni de l'addition $(a,b) \mapsto -\epsilon \log\left(e^{-a/\epsilon} + e^{-b/\epsilon}\right)$ et de la multiplication $(a,b) \mapsto a+b$. Pour tout $\epsilon > 0$, $\mathbb{R}_\epsilon$ est isomorphe au semi-corps usuel des réels positifs, $(\mathbb{R}_+, +, \times)$, mais pour $\epsilon = 0^+$, $\mathbb{R}_\epsilon$ n'est autre que le semi-anneau min-plus. Cette idée a été introduite par Maslov [156], motivé par l'étude des asymptotiques de type WKB d'équations de Schrödinger. Ce point de vue permet d'utiliser des résultats algébriques pour résoudre des problèmes d'asymptotiques, puisque les équations limites ont souvent un caractère min-plus linéaire.

Cette déformation apparaît classiquement en théorie des grandes déviations à la loi des grands nombres : dans ce contexte, les objets limites sont des mesures idempotentes au sens de Maslov. Voir [1], [170], [61], pour les relations entre l'algèbre max-plus et les grandes déviations, voir aussi [57], [56], [55] pour des applications de ces idées aux perturbations singulières de valeurs propres. La même déformation est à l'origine de nombreux travaux actuels en géométrie tropicale, à la suite de Viro [182].

English version

The role of min-plus algebra in asymptotic problems becomes obvious when writing Equations (7) when $\epsilon \to 0^+$. Formally, min-plus algebra may be thought of as the limit of a deformation of classical algebra, by introducing the semi-field $\mathbb{R}_\epsilon$, which is the set $\mathbb{R} \cup \{+\infty\}$, equipped with the addition $(a,b) \mapsto -\epsilon \log\left(e^{-a/\epsilon} + e^{-b/\epsilon}\right)$ and the multiplication $(a,b) \mapsto a+b$. For all $\epsilon > 0$, $\mathbb{R}_\epsilon$ is isomorphic to the semi-field of usual real positive numbers, $(\mathbb{R}_+, +, \times)$, but for $\epsilon = 0^+$, $\mathbb{R}_\epsilon$ coincides with the min-plus semiring. This idea was introduced by Maslov [156], motivated by the study of WKB-type asymptotics of Schrödinger equations. This point of view allows one to use algebraic results in asymptotics problems, since the limit equations have often some kind of min-plus linear structure.

This deformation appears classically in large deviation theory: in this context, the limiting objects are idempotent measures, in the sense of Maslov. See [1], [170], [61] for the relation between max-plus algebra and large deviations. See also [57], [56], [55] for the application of such ideas to singular perturbation problems for matrix eigenvalues. The same deformation is at the origin of many current works in tropical geometry, in the line initiated by Viro [182].

MC2 Team

3. Research Program

3.1. Introduction

We are mainly concerned with complex fluid mechanics problems. The complexity consists of the rheological nature of the fluids (non newtonian fluids), of the coupling phenomena (in shape optimization problems), of the geometry (micro-channels) or of multi-scale phenomena arising in turbulence or in tumor growth modeling. Our goal is to understand these phenomena and to simulate and/or to control them. The subject is wide and we will restrict ourselves to three directions: the first one consists in studying low Reynolds number interface problems in multi-fluid flows with applications to complex fluids, microfluidics and biology - the second one deals with numerical simulation of Newtonian fluid flows with emphasis on the coupling of methods to obtain fast solvers.

Even if we deal with several kinds of applications, there is a strong scientific core at each level of our project. Concerning the model, we are mainly concerned with incompressible flows and we work with the classical description of incompressible fluid dynamics. For the numerical methods, we use the penalization method to describe the obstacles or the boundary conditions for high Reynolds flows, for shape optimization, for interface problems in biology or in microfluidics. This allows us to use only cartesian meshes. Moreover, we use the level-set method for interface problems, for shape optimization and for fluid structure interaction. Finally, for the implementation, strong interaction exists between the members of the team and the modules of the numerical codes are used by all the team and we want to build the platform **eLYSe** to systematize this approach.

3.2. Multi-fluid flows and application for complex fluids, microfluidics

Participants: Angelo Iollo, Charles-Henri Bruneau, Thierry Colin, Mathieu Colin, Kévin Santugini.

Multi-fluid flows,microfluidics

By a complex fluid, we mean a fluid containing some mesoscopic objects, i.e. structures whose size is intermediate between the microscopic size and the macroscopic size of the experiment. The aim is to study complex fluids containing surfactants in large quantities. It modifies the viscosity properties of the fluids and surface-tension phenomena can become predominant.

Microfluidics is the study of fluids in very small quantities, in micro-channels (a micro-channel is typically 1 cm long with a section of 50 μ m$\times$ 50 μ m). They are many advantages of using such channels. First, one needs only a small quantity of liquid to analyze the phenomena. Furthermore, very stable flows and quite unusual regimes may be observed, which enables to perfom more accurate measurements. The idea is to couple numerical simulations with experiments to understand the phenomena, to predict the flows and compute some quantities like viscosity coefficients for example. Flows in micro-channels are often at low Reynolds numbers. The hydrodynamical part is therefore stable. However, the main problem is to produce real 3D simulations covering a large range of situations. For example we want to describe diphasic flows with surface tension and sometimes surface viscosity. Surface tension enforces the stability of the flow. The size of the channel implies that one can observe some very stable phenomena. For example, using a "T" junction, a very stable interface between two fluids can be observed. In a cross junction, one can also have formation of droplets that travel along the channel. Some numerical difficulties arise from the surface tension term. With an explicit discretization of this term, a restrictive stability condition appears for very slow flows [77]. Our partner is the LOF, a Rhodia-Bordeaux 1-CNRS laboratory.

One of the main points is the wetting phenomena at the boundary. Note that the boundary conditions are fundamental for the description of the flow since the channels are very shallow. The wetting properties cannot be neglected at all. Indeed, for the case of a two non-miscible fluids system, if one considers no-slip boundary conditions, then since the interface is driven by the velocity of the fluids, it shall not move on the boundary. The experiments shows that this is not true: the interface is moving and in fact all the dynamics start from the boundary and then propagate in the whole volume of fluids. Even with low Reynolds numbers, the wetting effects can induce instabilities and are responsible of hardly predictable flows. Moreover, the fluids that are used are often visco-elastic and exhibit "unusual" slip length. Therefore, we cannot use standard numerical codes and have to adapt the usual numerical methods to our case to take into account the specificities of our situations. In Johana Pinilla's thesis the Cox law has been implemented successfully to allow the interface to move properly between two Newtonian fluids of various viscosity or one Newtonian and one non-Newtonian fluid. Moreover, we want to obtain reliable models and simulations that can be as simple as possible and that can be used by our collaborators. As a summary, the main specific points of the physics are: the multi-fluid simulations at low Reynolds number, the wetting problems and the surface tension that are crucial, the 3D characteristic of the flows, the boundary conditions that are fundamental due to the size of the channels. We need to handle complex fluids. Our collaborators in this lab are H. Bordiguel, J.-B. Salmon, P. Guillot, A. Colin.

The evolution of non-newtonian flows in webs of micro-channels are therefore useful to understand the mixing of oil, water and polymer for enhanced oil recovery for example. Complex fluids arising in cosmetics are also of interest. We also need to handle mixing processes.

3.3. Cancer modeling

Participants: Sébastien Benzekry, Thierry Colin, Angelo Iollo, Clair Poignard, Olivier Saut, Lisl Weynans.

Tumor growth, cancer, metastasis

As in microfluidics, the growth of a tumor is a low Reynolds number flow. Several kinds of interfaces are present (membranes, several populations of cells,...) The biological nature of the tissues impose the use of different models in order to describe the evolution of tumor growth. The complexity of the geometry, of the rheological properties and the coupling with multi-scale phenomena is high but not far away from those encountered in microfluidics and the models and methods are close.

The challenge is twofold. On one hand, we wish to understand the complexity of the coupling effects between the different levels (cellular, genetic, organs, membranes, molecular). Trying to be exhaustive is of course hopeless, however it is possible numerically to isolate some parts of the evolution in order to better understand the interactions. Another strategy is to test *in silico* some therapeutic innovations. An example of such a test is given in [88] where the efficacy of radiotherapy is studied and in [89] where the effects of anti-invasive agents is investigated. It is therefore useful to model a tumor growth at several stage of evolution. The macroscopic continuous model is based on Darcy's law which seems to be a good approximation to describe the flow of the tumor cells in the extra-cellular matrix [54], [78], [79]. It is therefore possible to develop a two-dimensional model for the evolution of the cell densities. We formulate mathematically the evolution of the cell densities in the tissue as advection equations for a set of unknowns representing the density of cells with position (x, y) at time t in a given cycle phase. Assuming that all cells move with the same velocity given by Darcy's law and applying the principle of mass balance, one obtains the advection equations with a source term given by a cellular automaton. We assume diffusion for the oxygen and the diffusion constant depends on the density of the cells. The source of oxygen corresponds to the spatial location of blood vessels. The available quantities of oxygen interact with the proliferation rate given by the cellular automaton [88].

Another axis of these theoretical investigations is the study of several processes in cancer biology (with a major focus on metastasis) for applications in theoretical and experimental onco-biology as well as preclinical and clinical studies. This axis regroups several projects for which our approach can be decomposed into three steps. First, we base ourselves on a detailed study of the particular biological process, based on the available literature and in close collaboration with biologists and the available data. In a second step, we reduce the

biological dynamics to its more essential components and build mathematical models able to simulate the process, to address the particular biological question under investigation and to give nontrivial insights on the overall complex combination of these dynamics. Eventually, the last step consists in confronting the models to the data, using statistical parameter estimation methods, in order to identify theories or hypothesis that could or could not have generated the data and thus improve the biological understanding or identify optimal therapeutic strategies.

A forthcoming investigation in cancer treatment simulation is the influence of the electrochemotherapy [83] on the tumor growth. Electrochemotherapy consists in imposing to the malignant tumor high voltage electric pulses so that the plasma membrane of carcinoma cells is permeabilized. Biologically active molecules such as bleomycin, which usually cannot diffuse through the membrane, may then be internalized. A work in progress (C.Poignard [87] in collaboration with the CNRS lab of physical vectorology at the Institut Gustave Roussy) consists in modelling electromagnetic phenomena at the cell scale. A coupling between the microscopic description of the electroporation of cells and its influence on the global tumor growth at the macroscopic scale is expected. Another key point is the parametrization of the models in order to produce image-based simulations.

The second challenge is more ambitious. Mathematical models of cancer have been extensively developed with the aim of understanding and predicting tumor growth and the effects of treatments. In vivo modeling of tumors is limited by the amount of information available. However, in the last few years there have been dramatic increases in the range and quality of information available from non-invasive imaging methods, so that several potentially valuable imaging measurements are now available to quantitatively measure tumor growth, assess tumor status as well as anatomical or functional details. Using different methods such as the CT scan, magnetic resonance imaging (MRI), or positron emission tomography (PET), it is now possible to evaluate and define tumor status at different levels: physiological, molecular and cellular.

In this context, the present project aims at supporting the decision process of oncologists in the definition of therapeutic protocols via quantitative methods. The idea is to build mathematically and physically sound phenomenological models that can lead to patient-specific full-scale simulations, starting from data collected typically through medical imagery like CT scans, MRIs and PET scans or by quantitative molecular biology for leukemia. Our ambition is to provide medical doctors with patient-specific tumor growth models able to estimate, on the basis of previously collected data and within the limits of phenomenological models, the evolution at subsequent times of the pathology and possibly the response to the therapies.

The final goal is to provide numerical tools in order to help to answer to the crucial questions for a clinician:
When to start a treatment?
When to change a treatment?
When to stop a treatment?
Also we intend to incorporate real-time model information for improving the precision and effectiveness of non-invasive or micro-invasive tumor ablation techniques like acoustic hyperthermia, electroporation, radiofrequency or cryo-ablation.

We will specifically focus on the following pathologies: Lung and liver metastasis of a distant tumor
Low grade and high grade gliomas, meningiomas
Chronic myelogenous leukemia
These pathologies have been chosen because of the existing collaborations between the applied mathematics department of University of Bordeaux and the Institut Bergonié.

Our approach. Our approach is deterministic and spatial: it is based on solving an inverse problem based on imaging data. Models are of partial differential equation (PDE) type. They are coupled with a process of data assimilation based on imaging. We already have undertaken test cases on patients that are followed at Bergonié for lung metastases of thyroid tumors. These patients have a slowly evolving, asymptomatic metastatic disease, monitored by CT scans. On two thoracic images relative to successive times, the volume of the tumor under investigation is extracted by segmentation. To test our method, we chose patients without treatment and for whom we had at least three successive.

3.4. Newtonian fluid flows simulations and their analysis

Participants: Charles-Henri Bruneau, Angelo Iollo, Iraj Mortazavi, Michel Bergmann, Lisl Weynans.

Simulation, Analysis

It is very exciting to model complex phenomena for high Reynolds flows and to develop methods to compute the corresponding approximate solutions, however a well-understanding of the phenomena is necessary. Classical graphic tools give us the possibility to visualize some aspects of the solution at a given time and to even see in some way their evolution. Nevertheless in many situations it is not sufficient to understand the mechanisms that create such a behavior or to find the real properties of the flow. It is then necessary to carefully analyze the flow, for instance the vortex dynamics or to identify the coherent structures to better understand their impact on the whole flow behavior.
The various numerical methods used or developed to approximate the flows depend on the studied phenomenon. Our goal is to compute the most reliable method for each situation.
The first method, which is affordable in 2D, consists in a directly solving of the genuine Navier-Stokes equations in primitive variables (velocity-pressure) on Cartesian domains [64]. The bodies, around which the flow has to be computed are modeled using the penalization method (also named Brinkman-Navier-Stokes equations). This is an immersed boundary method in which the bodies are considered as porous media with a very small intrinsic permeability [55]. This method is very easy to handle as it consists only in adding a mass term U/K in the momentum equations. The boundary conditions imposed on artificial boundaries of the computational domains avoid any reflections when vortices cross the boundary. To make the approximation efficient enough in terms of CPU time, a multi-grid solver with a cell by cell Gauss-Seidel smoother is used.
The second type of methods is the vortex method. It is a Lagrangian technique that has been proposed as an alternative to more conventional grid-based methods. Its main feature is that the inertial nonlinear term in the flow equations is implicitly accounted for by the transport of particles. The method thus avoids to a large extent the classical stability/accuracy dilemma of finite-difference or finite-volume methods. This has been demonstrated in the context of computations for high Reynolds number laminar flows and for turbulent flows at moderate Reynolds numbers [72]. This method has recently enabled us to obtain new results concerning the three-dimensional dynamics of cylinder wakes.
The third method is to develop reduced order models (ROM) based on a Proper Orthogonal Decomposition (POD) [80]. The POD consists in approximating a given flow field $U(x,t)$ with the decomposition

$$U(x,t) = \sum_i a_i(t)\phi_i(x),$$

where the basis functions are empirical in the sense that they derive from an existing data base given for instance by one of the methods above. Then the approximation of Navier-Stokes equations for instance is reduced to solving a low-order dynamical system that is very cheap in terms of CPU time. Nevertheless the ROM can only restitute what is contained in the basis. Our challenge is to extend its application in order to make it an actual prediction tool.
The fourth method is a finite volume method on cartesian grids to simulate compressible Euler or Navier Stokes Flows in complex domains. An immersed boundary-like technique is developed to take into account boundary conditions around the obstacles with order two accuracy.

3.5. Flow control and shape optimization

Participants: Charles-Henri Bruneau, Angelo Iollo, Iraj Mortazavi, Michel Bergmann.

Flow Control, Shape Optimization

Flow simulations, optimal design and flow control have been developed these last years in order to solve real industrial problems : vortex trapping cavities with CIRA (Centro Italiano Ricerche Aerospaziali), reduction of vortex induced vibrations on deep sea riser pipes with IFP (Institut Français du Pétrole), drag reduction of a ground vehicle with Renault or in-flight icing with Bombardier and Pratt-Wittney are some examples of possible applications of these researches. Presently the recent creation of the competitiveness cluster on aeronautics, space and embedded systems (AESE) based also in Aquitaine provides the ideal environment to extend our applied researches to the local industrial context. There are two main streams: the first need is to produce direct numerical simulations, the second one is to establish reliable optimization procedures.
In the next subsections we will detail the tools we will base our work on, they can be divided into three points: to find the appropriate devices or actions to control the flow; to determine an effective system identification technique based on the trace of the solution on the boundary; to apply shape optimization and system identification tools to the solution of inverse problems found in object imaging and turbomachinery.

3.5.1. Control of flows

There are mainly two approaches: passive (using passive devices on some specific parts that modify the shear forces) or active (adding locally some energy to change the flow) control.

The passive control consists mainly in adding geometrical devices to modify the flow. One idea is to put a porous material between some parts of an obstacle and the flow in order to modify the shear forces in the boundary layer. This approach may pose remarkable difficulties in terms of numerical simulation since it would be necessary, a priori, to solve two models: one for the fluid, one for the porous medium. However, by using the penalization method it becomes a feasible task [60]. This approach has been now used in several contexts and in particular in the frame of a collaboration with IFP to reduce vortex induced vibrations [61]. Another technique we are interested in is to inject minimal amounts of polymers into hydrodynamic flows in order to stabilize the mechanisms which enhance hydrodynamic drag.

The active approach is addressed to conceive, implement and test automatic flow control and optimization aiming mainly at two applications : the control of unsteadiness and the control and optimization of coupled systems. Implementation of such ideas relies on several tools. The common challenges are infinite dimensional systems, Dirichlet boundary control, nonlinear tracking control, nonlinear partial state observation.

The bottom-line to obtain industrially relevant control devices is the energy budget. The energy required by the actuators should be less than the energy savings resulting from the control application. In this sense our research team has gained a certain experience in testing several control strategies with a doctoral thesis (E. Creusé) devoted to increasing the lift on a dihedral plane. Indeed the extension of these techniques to real world problems may reveal itself very delicate and special care will be devoted to implement numerical methods which permit on-line computing of actual practical applications. For instance the method can be successful to reduce the drag forces around a ground vehicle and a coupling with passive control is under consideration to improve the efficiency of each control strategy.

3.5.2. System identification

We remark that the problem of deriving an accurate estimation of the velocity field in an unsteady complex flow, starting from a limited number of measurements, is of great importance in many engineering applications. For instance, in the design of a feedback control, a knowledge of the velocity field is a fundamental element in deciding the appropriate actuator reaction to different flow conditions. In other applications it may be necessary or advisable to monitor the flow conditions in regions of space which are difficult to access or where probes cannot be fitted without causing interference problems.

The idea is to exploit ideas similar to those at the basis of the Kalman filter. The starting point is again a Galerkin representation of the velocity field in terms of empirical eigenfunctions. For a given flow, the POD modes can be computed once and for all based on Direct Numerical Simulation (DNS) or on highly resolved experimental velocity fields, such as those obtained by particle image velocimetry. An instantaneous velocity field can thus be reconstructed by estimating the coefficients $a_i(t)$ of its Galerkin representation. One simple approach to estimate the POD coefficients is to approximate the flow measurements in a least square sense, as in [76].

A similar procedure is also used in the estimation based on gappy POD, see [92] and [96]. However, these approaches encounter difficulties in giving accurate estimations when three-dimensional flows with complicated unsteady patterns are considered, or when a very limited number of sensors is available. Under these conditions, for instance, the least squares approach cited above (LSQ) rapidly becomes ill-conditioned. This simply reflects the fact that more and more different flow configurations correspond to the same set of measurements.

Our challenge is to propose an approach that combines a linear estimation of the coefficients $a_i(t)$ with an appropriate non-linear low-dimensional flow model, that can be readily implemented for real time applications.

3.5.3. Shape optimization and system identification tools applied to inverse problems found in object imaging and turbomachinery

We will consider two different objectives. The first is strictly linked to the level set methods that are developed for microfluidics. The main idea is to combine different technologies that are developed with our team: penalization methods, level sets, an optimization method that regardless of the model equation will be able to solve inverse or optimization problems in 2D or 3D. For this we have started a project that is detailed in the research program. See also [67] for a preliminary application.

As for shape optimization in aeronautics, the aeroacoustic optimization problem of propeller blades is addressed by means of an inverse problem and its adjoint equations. This problem is divided into three subtasks:

i) formulation of an inverse problem for the design of propeller blades and determination of the design parameters ii) derivation of an aeroacoustic model able to predict noise levels once the blade geometry and the flow field are given iii) development of an optimization procedure in order to minimize the noise emission by controlling the design parameters.

The main challenge in this field is to move from simplified models [81] to actual 3D model. The spirit is to complete the design performed with a simplified tool with a fully three dimensional inverse problem where the load distribution as well as the geometry of the leading edge are those provided by the meridional plane analysis [91]. A 3D code will be based on the compressible Euler equations and an immersed boundary technique over a cartesian mesh. The code will be implicit and parallel, in the same spirit as what was done for the meridional plane. Further development include the extension of the 3D immersed boundary approach to time-dependent phenomena. This step will allow the designer to take into account noise sources that are typical of internal flows. The task will consist in including time dependent forcing on the inlet and/or outlet boundary under the form of Fourier modes and in computing the linearized response of the system. The optimization will then be based on a direct approach, i.e., an approach where the control is the geometry of the boundary. The computation of the gradient is performed by an adjoint method, which will be a simple "byproduct" of the implicit solver. The load distribution as well as the leading edge geometry obtained by the meridional plane approach will be considered as constraints of the optimization, by projection of the gradient on the constraint tangent plane. These challenges will be undertaken in collaboration with Politecnico di Torino and EC Lyon.

MCTAO Project-Team

3. Research Program

3.1. Control Systems

Our effort is directed toward efficient methods for the *control* of real (physical) systems, based on a *model* of the system to be controlled. *System* refers to the physical plant or device, whereas *model* refers to a mathematical representation of it.

We mostly investigate nonlinear systems whose nonlinearities admit a strong structure derived from physics; the equations governing their behavior are then well known, and the modeling part consists in choosing what phenomena are to be retained in the model used for control design, the other phenomena being treated as perturbations; a more complete model may be used for simulations, for instance. We focus on systems that admit a reliable finite-dimensional model, in continuous time; this means that models are controlled ordinary differential equations, often nonlinear.

Choosing accurate models yet simple enough to allow control design is in itself a key issue; however, modeling or identification as a theory is not per se in the scope of our project.

The extreme generality and versatility of linear control do not contradict the often heard sentence "most real life systems are nonlinear". Indeed, for many control problems, a linear model is sufficient to capture the important features for control. The reason is that most control objectives are local, first order variations around an operating point or a trajectory are governed by a linear control model, and except in degenerate situations (non-controllability of this linear model), the local behavior of a nonlinear dynamic phenomenon is dictated by the behavior of first order variations. Linear control is the hard core of control theory and practice; it has been pushed to a high degree of achievement –see for instance some classics: [45], [35]– that leads to big successes in industrial applications (PID, Kalman filtering, frequency domain design, H^∞ robust control, etc...). It must be taught to future engineers, and it is still a topic of ongoing research.

Linear control by itself however reaches its limits in some important situations:

1. **Non local control objectives.** For instance, steering the system from a region to a reasonably remote other one (path planning and optimal control); in this case, local linear approximation cannot be sufficient.
   It is also the case when some domain of validity (e.g. stability) is prescribed and is larger than the region where the linear approximation is dominant.
2. **Local control at degenerate equilibria.** Linear control yields local stabilization of an equilibrium point based on the tangent linear approximation if the latter is controllable. When it is *not*, and this occurs in some physical systems at interesting operating points, linear control is irrelevant and specific nonlinear techniques have to be designed.
   This is in a sense an extreme case of the second paragraph in point 1 : the region where the linear approximation is dominant vanishes.
3. **Small controls.** In some situations, actuators only allow a very small magnitude of the effect of control compared to the effect of other phenomena. Then the behavior of the system without control plays a major role and we are again outside the scope of linear control methods.
4. **Local control around a trajectory.** Sometimes a trajectory has been selected (this appeals to point 1), and local regulation around this reference is to be performed. Linearization in general yields, when the trajectory is not a single equilibrium point, a *time-varying* linear system. Even if it is controllable, time-varying linear systems are not in the scope of most classical linear control methods, and it is better to incorporate this local regulation in the nonlinear design, all the more so as the linear approximation along optimal trajectories is, by nature, often non controllable.

Let us discuss in more details some specific problems that we are studying or plan to study: classification and structure of control systems in section 3.2 , optimal control, and its links with feedback, in section 3.3 , the problem of optimal transport in section 3.4 , and finally problems relevent to a specific class of systems where the control is "small" in section 3.5 .

3.2. Structure of nonlinear control systems

In most problems, choosing the proper coordinates, or the right quantities that describe a phenomenon, sheds light on a path to the solution. In control systems, it is often crucial to analyze the structure of the model, deduced from physical principles, of the plant to be controlled; this may lead to putting it via some transformations in a simpler form, or a form that is most suitable for control design. For instance, equivalence to a linear system may allow to use linear control; also, the so-called "flatness" property drastically simplifies path planning [40], [51].

A better understanding of the "set of nonlinear models", partly classifying them, has another motivation than facilitating control design for a given system and its model: it may also be a necessary step towards a theory of "nonlinear identification" and modeling. Linear identification is a mature area of control science; its success is mostly due to a very fine knowledge of the structure of the class of linear models: similarly, any progress in the understanding of the structure of the class of nonlinear models would be a contribution to a possible theory of nonlinear identification.

These topics are central in control theory, but raise very difficult mathematical questions: static feedback classification is a geometric problem which is feasible in principle, although describing invariants explicitly is technically very difficult; and conditions for dynamic feedback equivalence and linearization raise unsolved mathematical problems, that make one wonder about decidability [0].

3.3. Optimal control and feedback control, stabilization

3.3.1. Optimal control.

Mathematically speaking, optimal control is the modern branch of the calculus of variations, rather well established and mature [18], [49], [26], [58]. Relying on Hamiltonian dynamics is now prevalent, instead of the standard Lagrangian formalism of the calculus of variations. Also, coming from control engineering, constraints on the control (for instance the control is a force or a torque, which are naturally bounded) or the state (for example in the shuttle atmospheric re-entry problem there is a constraint on the thermal flux) are imposed; the ones on the state are usual but these on the state yield more complicated necessary optimality conditions and an increased intrinsic complexity of the optimal solutions. Also, in the modern treatment, ad-hoc numerical schemes have to be derived for effective computations of the optimal solutions.

What makes optimal control an applied field is the necessity of computing these optimal trajectories, or rather the controls that produce these trajectories (or, of course, close-by trajectories). Computing a given optimal trajectory and its control as a function of time is a demanding task, with non trivial numerical difficulties: roughly speaking, the Pontryagin Maximum Principle gives candidate optimal trajectories as solutions of a two point boundary value problem (for an ODE) which can be analyzed using mathematical tools from geometric control theory or solved numerically using shooting methods. Obtaining the *optimal synthesis* –the optimal control as a function of the state– is of course a more intricate problem [26], [31].

[0] Consider the simple system with state $(x, y, z) \in I\!R^3$ and two controls that reads $\dot{z} = (\dot{y} - z\dot{x})^2\dot{x}$ after elimination of the controls; it is not known whether it is equivalent to a linear system, or flat; this is because the property amounts to existence of a formula giving the general solution as a function of two arbitrary functions of time and their derivatives up to a certain order, but no bound on this order is known a priori, even for this very particular example.

These questions are not only academic for minimizing a cost is *very* relevant in many control engineering problems. However, modern engineering textbooks in nonlinear control systems like the "best-seller" [42] hardly mention optimal control, and rather put the emphasis on designing a feedback control, as regular and explicit as possible, satisfying some qualitative (and extremely important!) objectives: disturbance attenuation, decoupling, output regulation or stabilization. Optimal control is sometimes viewed as disconnected from automatic control... we shall come back to this unfortunate point.

3.3.2. Feedback, control Lyapunov functions, stabilization.

A control Lyapunov function **(CLF)** is a function that can be made a Lyapunov function (roughly speaking, a function that decreases along all trajectories, some call this an "artificial potential") for the closed-loop system corresponding to *some* feedback law. This can be translated into a partial differential relation sometimes called "Artstein's (in)equation" [21]. There is a definite parallel between a CLF for stabilization, solution of this differential inequation on the one hand, and the value function of an optimal control problem for the system, solution of a HJB equation on the other hand. Now, optimal control is a quantitative objective while stabilization is a qualitative objective; it is not surprising that Artstein (in)equation is very under-determined and has many more solutions than HJB equation, and that it may (although not always) even have smooth ones.

We have, in the team, a longstanding research record on the topic of construction of CLFs and stabilizing feedback controls. This is all the more interesting as our line of research has been pointing in almost opposite directions. [36], [55], [57] insist on the construction of continuous feedback, hence smooth CLFs whereas, on the contrary, [34], [59], [60] proceed with a very fine study of non-smooth CLFs, yet good enough (semi-concave) that they can produce a reasonable discontinuous feedback with reasonable properties.

3.4. Optimal Transport

We believe that matching optimal transport with geometric control theory is one originality of our team. We expect interactions in both ways.

The study of optimal mass transport problems in the Euclidean or Riemannian setting has a long history which goes from the pioneer works of Monge [53] and Kantorovitch [46] to the recent revival initiated by fundamental contributions due to Brenier [32] and McCann [52].

Th same transportation problems in the presence of differential constraints on the set of paths —like being an admissible trajectory for a control system— is quite new. The first contributors were Ambrosio and Rigot [19] who proved the existence and uniqueness of an optimal transport map for the Monge problem associated with the squared canonical sub-Riemannian distance on the Heisenberg groups. This result was extended later by Agrachev and Lee [16], then by Figalli and Rifford [37] who showed that the Ambrosio-Rigot theorem holds indeed true on many sub-Riemannian manifolds satisfying reasonable assumptions. The problem of existence and uniqueness of an optimal transport map for the squared sub-Riemannian distance on a general complete sub-Riemannian manifold remains open; it is strictly related to the regularity of the sub-Riemannian distance in the product space, and remains a formidable challenge. Generalized notions of Ricci curvatures (bounded from below) in metric spaces have been developed recently by Lott and Villani [50] and Sturm [63], [64]. A pioneer work by Juillet [43] captured the right notion of curvature for subriemannian metric in the Heisenberg group; Agrachev and Lee [17] have elaborated on this work to define new notions of curvatures in three dimensional sub-Riemannian structures. The optimal transport approach happened to be very fruitful in this context. Many things remain to do in a more general context.

3.5. Small controls and conservative systems, averaging

Using averaging techniques to study small perturbations of integrable Hamiltonian systems dates back to H. Poincaré or earlier; it gives an approximation of the (slow) evolution of quantities that are preserved in the non-perturbed system. It is very subtle in the case of multiple periods but more elementary in the single period case, here it boils down to taking the average of the perturbation along each periodic orbit; see for instance [20], [62].

When the "perturbation" is a control, these techniques may be used after deciding how the control will depend on time and state and other quantities, for instance it may be used after applying the Pontryagin Maximum Principle as in [23], [24], [33], [41]. Without deciding the control a priori, an "average control system" may be defined as in [22].

The focus is then on studying into details this simpler "averaged" problem, that can often be described by a Riemannian metric for quadratic costs or by a Finsler metric for costs like minimum time.

This line of research stemmed out of applications to space engineering, see section 4.1 . For orbit transfer in the two-body problem, an important contribution was made by B. Bonnard, J.-B. Caillau and J. Gergaud [24] in explicitly computing the solutions of the average system obtained after applying Pontryagin Maximum Principle to minimizing a quadratic integral cost; this yields an explicit calculation of the optimal control law itself. Studying the Finsler metric issued form the time-minimal case is in progress.

MEPHYSTO Team

3. Research Program

3.1. From statistical physics to continuum mechanics

Whereas numerical methods in nonlinear elasticity are well-developed and reliable, constitutive laws used for rubber in practice are phenomenological and generally not very precise. On the contrary, at the scale of the polymer-chain network, the physics of rubber is very precisely described by statistical physics. The main challenge in this field is to understand how to derive macroscopic constitutive laws for rubber-like materials from statistical physics.

At the continuum level, rubber is modelled by an energy E defined as the integral over a domain D of $\mathbb{R}^d$ of some energy density W depending only locally on the gradient of the deformation u: $E(u) = \int_D W(\nabla u(x))dx$. At the microscopic level (say 100nm), rubber is a network of cross-linked and entangled polymer chains (each chain is made of a sequence of monomers). At this scale the physics of polymer chains is well-understood in terms of statistical mechanics: monomers thermally fluctuate according to the Boltzmann distribution [46]. The associated Hamiltonian of a network is typically given by a contribution of the polymer chains (using self-avoiding random bridges) and a contribution due to steric effects (rubber is packed and monomers are surrounded by an excluded volume). The main challenge is to understand how this statistical physics picture yields rubber elasticity. Treloar assumed in [56] that for a piece of rubber undergoing some macroscopic deformation, the cross-links do not fluctuate and follow the macroscopic deformation, whereas between two cross-links, the chains fluctuate. This is the so-called affine assumption. Treloar's model is in rather good agreement with mechanical experiments in small deformation. In large deformation however, it overestimates the stress. A natural possibility to relax Treloar's model consists in relaxing the affine assumption while keeping the network description, which allows one to distinguish between different rubbers. This can be done by assuming that the deformation of the cross-links minimizes the free energy of the polymer chains, the deformation being fixed at the boundary of the macroscopic domain D. This gives rise to a "variational model". The analysis of the asymptotic behavior of this model as the typical length of a polymer chain vanishes has the same flavor as the homogenization theory of integral functionals in nonlinear elasticity (see [41], [52] in the periodic setting, and [42] in the random setting).

Our aim is to relate qualitatively and quantitatively the (precise but unpractical) statistical physics picture to explicit macroscopic constitutive laws that can be used for practical purposes.

In collaboration with R. Alicandro (Univ. Cassino, Italy) and M. Cicalese (Univ. Munich, Germany), A. Gloria analyzed in [1] the (asymptotic) Γ-convergence of the variational model for rubber, in the case when the polymer chain network is represented by some ergodic random graph. The easiest such graph is the Delaunay tessellation of a point set generated as follows: random hard spheres of some given radius ρ are picked randomly until the domain is jammed (the so-called random parking measure of intensity ρ). With M. Penrose (Univ. Bath, UK), A. Gloria studied this random graph in this framework [6]. With P. Le Tallec (Mechanics department, Ecole polytechnique, France), M. Vidrascu (project-team REO, Inria Paris-Rocquencourt), and A. Gloria introduced and tested in [15] a numerical algorithm to approximate the homogenized energy density, and observed that this model compares well to rubber elasticity qualitatively.

These preliminary results show that the variational model has the potential to explain qualitatively and quantitatively how rubber elasticity emerges from polymer physics. In order to go further and obtain more quantitative results and rigorously justify the model, we have to address several questions of analysis, modelling, scientific computing, inverse problems, and physics.

3.2. Quantitative stochastic homogenization

Whereas the approximation of homogenized coefficients is an easy task in periodic homogenization, this is a highly nontrivial task for stochastic coefficients. This is in order to analyze numerical approximation methods of the homogenized coefficients that F. Otto (MPI for mathematics in the sciences, Leipzig, Germany) and A. Gloria obtained the first quantitative results in stochastic homogenization [4]. The development of a complete stochastic homogenization theory seems to be ripe for the analysis and constitutes the second major objective of this section.

In order to develop a quantitative theory of stochastic homogenization, one needs to quantitatively understand the corrector equation (3). Provided A is stationary and ergodic, it is known that there exists a unique random field ϕ_ξ which is a distributional solution of (3) almost surely, such that $\nabla\phi_\xi$ is a stationary random field with bounded second moment $\langle|\nabla\phi_\xi|^2\rangle < \infty$, and with $\phi(0) = 0$. Soft arguments do not allow to prove that ϕ_ξ may be chosen stationary (this is wrong in dimension $d = 1$). In [4], [5] F. Otto and A. Gloria proved that, in the case of discrete elliptic equations with iid conductances, there exists a unique stationary corrector ϕ_ξ with vanishing expectation in dimension $d > 2$. Although it cannot be bounded, it has bounded finite moments of any order:

$$\langle|\phi_\xi|^q\rangle < \infty \text{ for all } q \geq 1. \tag{53}$$

They also proved that the variance of spatial averages of the energy density $(\xi + \nabla\phi_\xi) \cdot A(\xi + \nabla\phi_\xi)$ on balls of radius R decays at the rate R^{-d} of the central limit theorem. These are the *first optimal quantitative results* in stochastic homogenization.

The proof of these results, which is inspired by [53], is based on the insight that coefficients such as the Poisson random inclusions are special in the sense that the associated probability measure satisfies a spectral gap estimate. Combined with elliptic regularity theory, this spectral gap estimate quantifies ergodicity in stochastic homogenization. This systematic use of tools from statistical physics has opened the way to the quantitative study of stochastic homogenization problems, which we plan to fully develop.

3.3. Nonlinear Schrödinger equations

As well known, the (non)linear Schrödinger equation

$$\partial_t\varphi(t,x) = -\Delta\varphi(t,x) + \lambda V(x)\varphi(t,x) + g|\varphi|^2\varphi(t,x), \quad \varphi(0,x) = \varphi_0(x) \tag{54}$$

with coupling constants $g \in \mathbb{R}, \lambda \in \mathbb{R}_+$ and real potential V (possibly depending also on time) models many phenomena of physics.

When in the equation (5) above one sets $\lambda = 0, g \neq 0$, one obtains the nonlinear (focusing of defocusing) Schrödinger equation. It is used to model light propagation in optical fibers. In fact, it then takes the following form:

$$i\partial_z\varphi(t,z) = -\beta(z)\partial_t^2\varphi(t,z) + \gamma(z)|\varphi(t,z)|^2\varphi(z,t), \tag{55}$$

where β and γ are functions that characterize the physical properties of the fiber, t is time and z the position along the fiber. Several issues are of importance here. Two that will be investigated within the MEPHYSTO project are: the influence of a periodic modulation of the fiber parameters β and γ and the generation of so-called “rogue waves” (which are solutions of unusually high amplitude) in such systems.

If $g = 0, \lambda \neq 0$, V is a random potential, and φ_0 is deterministic, this is the standard random Schrödinger equation describing for example the motion of an electron in a random medium. The main issue in this setting is the determination of the regime of Anderson localization, a property characterized by the boundedness in time of the second moment $\int x^2 |\varphi(t,x)|^2 dx$ of the solution. If this second moment remains bounded in time, the solution is said to be localized. Whereas it is known that the solution is localized in one dimension for all (suitable) initial data, both localized and delocalized solutions exist in dimension 3 and it remains a major open problem today to prove this, cf. [44].

If now $g \neq 0, \lambda \neq 0$ and V is still random, but $|g| \ll \lambda$, a natural question is whether, and in which regime, one-dimensional Anderson localization perdures. Indeed, Anderson localization can be affected by the presence of the nonlinearity, which corresponds to an interaction between the electrons or atoms. Much numerical and some analytical work has been done on this issue (see for example [47] for a recent work at PhLAM, Laser physics department, Univ. Lille 1), but many questions remain, notably on the dependence of the result on the initial conditions, which, in a nonlinear system, may be very complex. The cold atoms team of PhLAM (Garreau-Szriftgiser) is currently setting up an experiment to analyze the effect of the interactions in a Bose-Einstein condensate on a closely related localization phenomenon called "dynamical localization", in the kicked rotor, see below.

3.4. Dynamical localization and kicked rotors

The kicked rotor is a unitary discrete time dynamics proposed in the seventies in the context of studies on quantum chaos, and used recently as a "quantum simulator" for the Anderson model. It is a quantum equivalent of the standard map and is obtained by integrating a time-dependent linear Schrödinger equation with a time-periodic, very singular (delta comb) potential. It continues to pose considerable mathematical challenges, in particular the so-called "quantum suppression of classical chaos" in the presence of a strong potential, which remains an open problem from the mathematical point of view. It can be rephrased as follows: show that the H^1 norm of the solution is uniformly bounded in time (see [36] for more background). In more recent years, the question has arisen how the behavior of this system would change in the presence of a nonlinear term in the Schrödinger equation.

This problem displays both numerical and analytical challenges, in particular because of the difficulty to obtain long time simulations of the system and because of the presence of instabilities due to the nonlinearity. Preliminary theoretical results motivate some conjectures on the behavior of these systems, that we plan to validate empirically in a first step. Indeed, reliable long-time simulations of the system should allow us to get more insight into the behavior of the exact solutions in the unstable cases. One of the main difficulties for the numerical simulation is the intrinsic instability of the system, which magnifies quite rapidly the numerical error due to machine precision. This requires the use of multiprecision techniques in order to handle reasonably long times, even for moderate nonlinearities, and of the transparent boundary conditions recently introduced by members of the former SIMPAF project-team.

MISTIS Project-Team

3. Research Program

3.1. Mixture models

Participants: Angelika Studeny, Thomas Vincent, Alexis Arnaud, Jean-Baptiste Durand, Florence Forbes, Aina Frau Pascual, Alessandro Chiancone, Stéphane Girard, Marie-José Martinez.

Key-words: mixture of distributions, EM algorithm, missing data, conditional independence, statistical pattern recognition, clustering, unsupervised and partially supervised learning.

In a first approach, we consider statistical parametric models, θ being the parameter, possibly multi-dimensional, usually unknown and to be estimated. We consider cases where the data naturally divides into observed data $y = y_1, ..., y_n$ and unobserved or missing data $z = z_1, ..., z_n$. The missing data z_i represents for instance the memberships of one of a set of K alternative categories. The distribution of an observed y_i can be written as a finite mixture of distributions,

$$f(y_i \mid \theta) = \sum_{k=1}^{K} P(z_i = k \mid \theta) f(y_i \mid z_i, \theta) \ . \tag{56}$$

These models are interesting in that they may point out hidden variable responsible for most of the observed variability and so that the observed variables are *conditionally* independent. Their estimation is often difficult due to the missing data. The Expectation-Maximization (EM) algorithm is a general and now standard approach to maximization of the likelihood in missing data problems. It provides parameter estimation but also values for missing data.

Mixture models correspond to independent z_i's. They have been increasingly used in statistical pattern recognition. They enable a formal (model-based) approach to (unsupervised) clustering.

3.2. Markov models

Participants: Angelika Studeny, Thomas Vincent, Jean-Baptiste Durand, Florence Forbes.

Key-words: graphical models, Markov properties, hidden Markov models, clustering, missing data, mixture of distributions, EM algorithm, image analysis, Bayesian inference.

Graphical modelling provides a diagrammatic representation of the dependency structure of a joint probability distribution, in the form of a network or graph depicting the local relations among variables. The graph can have directed or undirected links or edges between the nodes, which represent the individual variables. Associated with the graph are various Markov properties that specify how the graph encodes conditional independence assumptions.

It is the conditional independence assumptions that give graphical models their fundamental modular structure, enabling computation of globally interesting quantities from local specifications. In this way graphical models form an essential basis for our methodologies based on structures.

The graphs can be either directed, e.g. Bayesian Networks, or undirected, e.g. Markov Random Fields. The specificity of Markovian models is that the dependencies between the nodes are limited to the nearest neighbor nodes. The neighborhood definition can vary and be adapted to the problem of interest. When parts of the variables (nodes) are not observed or missing, we refer to these models as Hidden Markov Models (HMM). Hidden Markov chains or hidden Markov fields correspond to cases where the z_i's in (1) are distributed according to a Markov chain or a Markov field. They are a natural extension of mixture models. They are widely used in signal processing (speech recognition, genome sequence analysis) and in image processing (remote sensing, MRI, etc.). Such models are very flexible in practice and can naturally account for the phenomena to be studied.

Hidden Markov models are very useful in modelling spatial dependencies but these dependencies and the possible existence of hidden variables are also responsible for a typically large amount of computation. It follows that the statistical analysis may not be straightforward. Typical issues are related to the neighborhood structure to be chosen when not dictated by the context and the possible high dimensionality of the observations. This also requires a good understanding of the role of each parameter and methods to tune them depending on the goal in mind. Regarding estimation algorithms, they correspond to an energy minimization problem which is NP-hard and usually performed through approximation. We focus on a certain type of methods based on variational approximations and propose effective algorithms which show good performance in practice and for which we also study theoretical properties. We also propose some tools for model selection. Eventually we investigate ways to extend the standard Hidden Markov Field model to increase its modelling power.

3.3. Functional Inference, semi- and non-parametric methods

Participants: Farida Enikeeva, Alessandro Chiancone, Stéphane Girard, Gildas Mazo, Seydou-Nourou Sylla, Pablo Mesejo Santiago.

Key-words: dimension reduction, extreme value analysis, functional estimation.

We also consider methods which do not assume a parametric model. The approaches are non-parametric in the sense that they do not require the assumption of a prior model on the unknown quantities. This property is important since, for image applications for instance, it is very difficult to introduce sufficiently general parametric models because of the wide variety of image contents. Projection methods are then a way to decompose the unknown quantity on a set of functions (*e.g.* wavelets). Kernel methods which rely on smoothing the data using a set of kernels (usually probability distributions) are other examples. Relationships exist between these methods and learning techniques using Support Vector Machine (SVM) as this appears in the context of *level-sets estimation* (see section 3.3.2). Such non-parametric methods have become the cornerstone when dealing with functional data [71]. This is the case, for instance, when observations are curves. They enable us to model the data without a discretization step. More generally, these techniques are of great use for *dimension reduction* purposes (section 3.3.3). They enable reduction of the dimension of the functional or multivariate data without assumptions on the observations distribution. Semi-parametric methods refer to methods that include both parametric and non-parametric aspects. Examples include the Sliced Inverse Regression (SIR) method [74] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [69], which is based on the modelling of distribution tails (see section 3.3.1). It differs from traditional statistics which focuses on the central part of distributions, *i.e.* on the most probable events. Extreme value theory shows that distribution tails can be modelled by both a functional part and a real parameter, the extreme value index.

3.3.1. Modelling extremal events

Extreme value theory is a branch of statistics dealing with the extreme deviations from the bulk of probability distributions. More specifically, it focuses on the limiting distributions for the minimum or the maximum of a large collection of random observations from the same arbitrary distribution. Let $X_{1,n} \leq ... \leq X_{n,n}$ denote n ordered observations from a random variable X representing some quantity of interest. A p_n-quantile of X is the value x_{p_n} such that the probability that X is greater than x_{p_n} is p_n, *i.e.*$P(X > x_{p_n}) = p_n$. When $p_n < 1/n$, such a quantile is said to be extreme since it is usually greater than the maximum observation $X_{n,n}$ (see Figure 1).

To estimate such quantiles therefore requires dedicated methods to extrapolate information beyond the observed values of X. Those methods are based on Extreme value theory. This kind of issue appeared in hydrology. One objective was to assess risk for highly unusual events, such as 100-year floods, starting from flows measured over 50 years. To this end, semi-parametric models of the tail are considered:

$$P(X > x) = x^{-1/\theta}\ell(x),\ x > x_0 > 0, \tag{57}$$

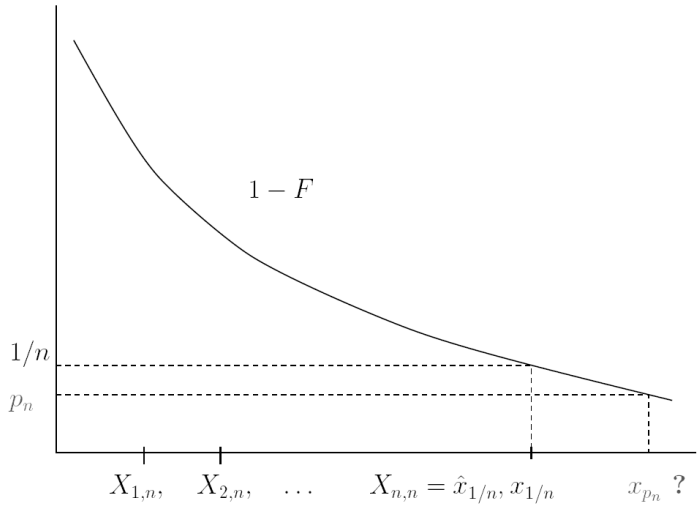


Figure 1. The curve represents the survival function $x \to P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\widehat{x}_{1/n} = X_{n,n}$. As illustrated in the figure, to estimate p_n-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.

where both the extreme-value index $\theta > 0$ and the function $\ell(x)$ are unknown. The function ℓ is a slowly varying function *i.e.* such that

$$\frac{\ell(tx)}{\ell(x)} \to 1 \text{ as } x \to \infty \tag{58}$$

for all $t > 0$. The function $\ell(x)$ acts as a nuisance parameter which yields a bias in the classical extreme-value estimators developed so far. Such models are often referred to as heavy-tail models since the probability of extreme events decreases at a polynomial rate to zero. It may be necessary to refine the model (2 ,3) by specifying a precise rate of convergence in (3). To this end, a second order condition is introduced involving an additional parameter $\rho \leq 0$. The larger ρ is, the slower the convergence in (3) and the more difficult the estimation of extreme quantiles.
More generally, the problems that we address are part of the risk management theory. For instance, in reliability, the distributions of interest are included in a semi-parametric family whose tails are decreasing exponentially fast. These so-called Weibull-tail distributions [9] are defined by their survival distribution function:

$$P(X > x) = \exp\{-x^{\theta}\ell(x)\},\ x > x_0 > 0. \tag{59}$$

Gaussian, gamma, exponential and Weibull distributions, among others, are included in this family. An important part of our work consists in establishing links between models (2) and (4) in order to propose new estimation methods. We also consider the case where the observations were recorded with a covariate information. In this case, the extreme-value index and the p_n-quantile are functions of the covariate. We propose estimators of these functions by using moving window approaches, nearest neighbor methods, or kernel estimators.

3.3.2. Level sets estimation

Level sets estimation is a recurrent problem in statistics which is linked to outlier detection. In biology, one is interested in estimating reference curves, that is to say curves which bound 90% (for example) of the population. Points outside this bound are considered as outliers compared to the reference population. Level sets estimation can be looked at as a conditional quantile estimation problem which benefits from a non-parametric statistical framework. In particular, boundary estimation, arising in image segmentation as well as in supervised learning, is interpreted as an extreme level set estimation problem. Level sets estimation can also be formulated as a linear programming problem. In this context, estimates are sparse since they involve only a small fraction of the dataset, called the set of support vectors.

3.3.3. Dimension reduction

Our work on high dimensional data requires that we face the curse of dimensionality phenomenon. Indeed, the modelling of high dimensional data requires complex models and thus the estimation of high number of parameters compared to the sample size. In this framework, dimension reduction methods aim at replacing the original variables by a small number of linear combinations with as small as a possible loss of information. Principal Component Analysis (PCA) is the most widely used method to reduce dimension in data. However, standard linear PCA can be quite inefficient on image data where even simple image distorsions can lead to highly non-linear data. Two directions are investigated. First, non-linear PCAs can be proposed, leading to semi-parametric dimension reduction methods [72]. Another field of investigation is to take into account the application goal in the dimension reduction step. One of our approaches is therefore to develop new Gaussian models of high dimensional data for parametric inference [67]. Such models can then be used in a Mixtures or Markov framework for classification purposes. Another approach consists in combining dimension reduction, regularization techniques, and regression techniques to improve the Sliced Inverse Regression method [74].

MODAL Project-Team

3. Research Program

3.1. Generative model design

The first objective of MODAL consists in designing, analyzing, estimating and evaluating new generative parametric models for multivariate and/or heterogeneous data. It corresponds typically to continuous and categorical data but it includes also other widespread ones like ordinal, functional, ranks,...Designed models have to take into account potential correlations between variables while being (1) justifiable and realistic, (2) meaningful and parsimoniously parameterized, (3) of low computational complexity. The main purpose is to identify a few theoretical and general principles for model generation, loosely dependent on the variable nature. In this context, we propose two concurrent approaches which could be general enough for dealing with correlation between many types of homogeneous or heterogeneous variables:

- Designs general models by combining two extreme models (full dependent and full independent) which are well-defined for most of variables;
- Uses kernels as a general way for dealing with multivariate and heterogeneous variables.

3.2. Data visualization

The second objective of MODAL is to propose meaningful and quite accurate low dimensional visualizations of data typically in two-dimensional (2D) spaces, less frequently in one-dimensional (1D) or three-dimensional (3D) spaces, by using the generative models designed in the first objective. We propose also to visualize simultaneously the data and the model. All visualizations will depend on the aim at hand (typically clustering, classification or density estimation). The main originality of this objective lies in the use of models for visualization, a strategy from which we expect to have a better control on the subjectivity necessarily induced by any graphical display. In addition, the proposed approach has to be general enough to be independent on the variable nature. Note that the visualization objective is consistent with the dissemination of our methodologies through specific softwares. Indeed, displaying data is an important step in the data analysis process.

MOKAPLAN Team

3. Research Program

3.1. Context

Optimal Mass Transportation is a mathematical research topic which started two centuries ago with Monge's work on "des remblais et déblais". This engineering problem consists in minimizing the transport cost between two given mass densities. In the 40's, Kantorovich [64] solved the dual problem and interpreted it as an economic equilibrium. The *Monge-Kantorovich* problem became a specialized research topic in optimization and Kantorovich obtained the 1975 Nobel prize in economics for his contributions to resource allocations problems. Following the seminal discoveries of Brenier in the 90's [35], Optimal Transportation has received renewed attention from mathematical analysts and the Fields Medal awarded in 2010 to C. Villani, who gave important contributions to Optimal Transportation and wrote the modern reference monograph [84], arrived at a culminating moment for this theory. Optimal Mass Transportation is today a mature area of mathematical analysis with a constantly growing range of applications (see below).

In the modern Optimal Mass Transportation problem, we are given two probability measures or "mass" densities : $d\rho_i(x_i)(= \rho_i(x_i)\, dx_i),\ i = 0, 1$ such that $\rho_i \geq 0,\ \int_{X_0} \rho_0(x_0)dx_0 = \int_{X_1} \rho_1(x_1)dx_1 = 1$, $X_i \subset R^n$. They are often referred to, respectively, source and target densities, support or spaces. The problem is the minimization of a *transportation cost*, $\mathcal{I}(M) = \int_{X_0} c(x, M(x))\, \rho_0(x)dx$ where c is a displacement *ground cost*, over all *volume preserving maps*$M \in \mathcal{MM} = \{M : X_0 \to X_1,\ M_{\#}d\rho_0 = d\rho_1\}$. Assuming that M is a diffeomorphism, this is equivalent to the *Jacobian equation*$det(DM(x))\rho_1(M(x)) = \rho_0(x)$. Most of the modern Optimal Mass Transportation theory has been developed for the Euclidean distance squared cost $c(x, y) = \|x - y)\|^2$ while the historic monge cost was the simple distance $c(x, y) = \|x - y\|$.

In the Euclidean distance squared ground cost, the problem is well posed and in the seminal work of Brenier [36], the optimal map is characterized as the gradient of a convex potential ϕ^* : $\mathcal{I}(\nabla\phi^*(x)) = \min_{M\in\mathcal{M}} \mathcal{I}(M)$. A formal substitution in the Jacobian equation gives the Monge-Ampère equation $det(D^2\phi^*)\rho_1(\nabla\phi^*(x)) = \rho_0(x)$ complemented by the *second boundary value* condition $\nabla\phi^*(X_0) \subset X_1$. Caffarelli [41] used this result to extend the regularity theory for the Monge-Ampère equation. He noticed in particular that Optimal Mass Transportation solutions, now called *Brenier solutions*, may have discontinuous gradients when the target density support X_1 is non convex and are therefore weaker than the Monge-Ampère potentials associated to Alexandrov measures (see [60] for a review of the different notions of Monge-Ampère solutions). The value function $\sqrt{\mathcal{I}(\nabla\phi^*)}$ is also known to be the *Wasserstein distance*$W_2(\rho_0, \rho_1)$ on the space of probability densities, see [84]. The *Computational Fluid Dynamic* formulation proposed by Brenier and Benamou in [2] introduces a time extension of the domain and leads to a convex but non smooth optimization problem : $\mathcal{I}(\nabla\phi^*) = \min_{(\rho,V)\in\mathcal{C}} \int_0^1 \int_X \frac{1}{2}\rho(t, x)\, \|V(t, x)\|^2\, dx\, dt$. with constraints : $\mathcal{C} = \{(\rho, V),\ \text{s.t}\ \partial_t\rho + div(\rho\, V) = 0,\ \rho(\{0, 1\}, .) = \rho_{\{0,1\}}(.)\}$. The time curves $t \to \rho(t, .)$ are geodesics between ρ_0 and ρ_1 for the Wasserstein distance. This formulation is a limit case of *Mean Fields games* [65], a large class of economic models introduced by Lasry and Lions. The Wasserstein distance and its connection to Optimal Mass Transportation also appears in the construction of semi-discrete Gradient Flows. This notion known as *JKO gradient flows* after its authors in [62] is a popular tool to study non-linear diffusion equations : the implicit Euler scheme $\rho_{k+1}^{dt} = argmin_{\rho(.)}\, F(\rho(.)) + \frac{1}{2\,dt}W_2(\rho(.), \rho_k^{dt})^2$ can be shown to converge $\rho_k^{dt}(.)) \to \rho^*(t, .)$ as $dt \to 0$ to the solution of the non linear continuity equation $\partial_t\rho^* + div(\rho^*\, \nabla(-\frac{\partial F}{\partial\rho}(\rho^*))) = 0,\ \rho^*(0, .) = \rho_0^{dt}(.)$. The prototypical example is given by $F(\rho) = \int_X \rho(x)\, \log(\rho(x)) + \rho(x)\, V(x)\, dx$ which corresponds to the classical Fokker-Planck equation. Extensions of the ground cost c have been actively studied recently, some are mentioned in the application section. Technical results culminating with the *Ma-Trudinger-Wang* condition [68] which gives necessary condition on c for the regularity of the solution of the Optimal Mass Transportation problem. More recently attention has risen on multi marginal Optimal Mass Transportation [59] and has been systematically studied

in [76] [79] [77] [78]. The data consists in an arbitrary (and even infinite) number N of densities (the marginals) and the ground cost is defined on a product space $c(x_0, x_1, ..., x_{n-1})$ of the same dimension. Several interesting applications belong to this class of models (see below).
Our focus is on numerical methods in Optimal Mass Transportation and applications. The simplest way to build a numerical method is to consider sum of dirac masses $\rho_0 = \sum_{i=1}^N \delta_{A_i} \quad \rho_1 = \sum_{j=1}^N \delta_{B_j}$. In that case the Optimal Mass Transportation problem reduces to combinatorial optimisation *assigment problem* between the points $\{A_i\}$s and $\{B_i\}$s : $\min_{\sigma \in Permut(1,N)} \frac{1}{N} \sum_{i=1}^N C_{i,\sigma(i)} C_{i,j} = \|A_i - B_j\|^2$. The complexity of the best (Hungarian or Auction) algorithm, see [33] for example, is $O(N^{\frac{5}{2}})$. An interesting variant is obtained when only the target measure is discrete. For instance $X_0 = \{\|x\| < 1\}$, $\rho_0 = \frac{1}{|X_0|} \quad \rho_1 = \frac{1}{N} \sum_{j=1}^N \delta_{y_j}$. It corresponds to the notion of Pogorelov solutions of the Monge-Ampère equation [80] and is also linked to Minkowski problem [31]. The optimal map is piecewise constant and the slopes are known. More precisely there exists N polygonal cells C_j such that $X_0 = \cup_j C_j$, $|C_j| = \frac{1}{N}$ and $\nabla \phi^*|_{C_j} = y_j$. Pogorelov proposed a constructive algorithm to build these solutions which has been refined and extended in particular in [50] [74] [72] [71]. The complexity is still not linear : $O(N^2 logN)$.
For general densities data, the original optimization problem is not tractable because of the volume preserving constraint on the map. Kantorovich dual formulation is a linear program but with a large number of constraints set over the product of the source and target space $X_0 \times X_1$. The CFD formulation [2]. preserves the convexity of the objective function and transforms the volume preserving constraint into a linear continuity equation (using a change of variable). We obtained a convex but non smooth optimization problem solved using an Augmented Lagrangian method [53], as originally proposed in [2]. It has been reinterpreted recently in the framework of proximal algorithms [75]. This approach is robust and versatile and has been reimplemented many times. It remains a first order optimization method and converges slowly. The cost is also increased by the additional artificial time dimension. An empirical complexity is $O(N^3 LogN)$ where N is the space discretization of the density. Several variants and extension of these methods have been implemented, in particular in [39] [30]. It is the only provably convergent method to compute Brenier (non C^1) solutions.
When interested in slightly more regular solutions which correspond to the assumption that the target support is convex, the recent *wide stencil* monotone finite difference scheme for the Monge-Ampère equation [55] can be adapted to the Optimal Mass Transportation problem. This is the topic of [7]. This approach is extremely fast as a Newton algorithm can be used to solve the discrete system. Numerical studies confirm this with a linear empirical complexity.

For other costs, JKO schemes, multi marginal extensions, partial transport ... efficient numerical methods are to be invented.

NACHOS Project-Team

3. Research Program

3.1. Scientific foundations

The teams focuses on physical applications dealing with electromagnetic or elastodynamic wave propagation in interaction with heterogeneous media and irregularly shaped structures. The underlying wave propagation phenomena can be purely unsteady or they can be periodic (because the imposed source term follows a time-harmonic evolution). In this context, the research activities undertaken by the team aim at developing innovative numerical methodologies putting the emphasis on several features:

- **Accuracy**. The foreseen numerical methods should rely on discretization techniques that best fit to the geometrical characteristics of the problems at hand. Methods based on unstructured, locally refined, even non-conforming, simplicial meshes are particularly attractive in this regard. In addition, the proposed numerical methods should also be capable to accurately describe the underlying physical phenomena that may involve highly variable space and time scales. Both objectives are generally addressed by studying so-called hp-adaptive solution strategies which combine h-adaptivity using local refinement/coarsening of the mesh and p-adaptivity using adaptive local variation of the interpolation order for approximating the solution variables. However, for physical problems involving strongly heterogeneous or high contrast propagation media, such a solution strategy may not be sufficient. Then, for dealing accurately with these situations, one has to design numerical methods that specifically address the multiscale nature of the underlying physical phenomena.
- **Numerical efficiency**. The simulation of unsteady problems most often relies on explicit time integration schemes. Such schemes are constrained by a stability criterion, linking some space and time discretization parameters, that can be very restrictive when the underlying mesh is highly non-uniform (especially for locally refined meshes). For realistic 3d problems, this can represent a severe limitation with regards to the overall computing time. One possible overcoming solution consists in resorting to an implicit time scheme in regions of the computational domain where the underlying mesh size is very small, while an explicit time scheme is applied elsewhere in the computational domain. The resulting hybrid explicit-implicit time integration strategy raises several challenging questions concerning both the mathematical analysis (stability and accuracy, especially for what concern numerical dispersion), and the computer implementation on modern high performance systems (data structures, parallel computing aspects). A second, often considered approach is to devise a local time strategy in the context of a fully explicit time integration scheme. Beside, when considering time-harmonic wave propagation problems, numerical efficiency is mainly linked to the solution of the system of algebraic equations resulting from the discretization in space of the underlying PDE model. Various strategies exist ranging from the more robust and efficient sparse direct solvers to the more flexible and cheaper (in terms of memory resources) iterative methods. Current trends tend to show that the ideal candidate will be a judicious mix of both approaches by relying on domain decomposition principles.
- **Computational efficiency**. Realistic 3d wave propagation problems involve the processing of very large volumes of data. The latter results from two combined parameters: the size of the mesh i.e the number of mesh elements, and the number of degrees of freedom per mesh element which is itself linked to the degree of interpolation and to the number of physical variables (for systems of partial differential equations). Hence, numerical methods must be adapted to the characteristics of modern parallel computing platforms taking into account their hierarchical nature (e.g multiple processors and multiple core systems with complex cache and memory hierarchies). In addition, appropriate parallelization strategies need to be designed that combine SIMD and MIMD programming paradigms.

From the methodological point of view, the research activities of the team are concerned with four main topics: (1) high order finite element type methods on unstructured or hybrid structured/unstructured meshes for the discretization of the considered systems of PDEs, (2) efficient time integration strategies for dealing with grid induced stiffness when using non-uniform (locally refined) meshes, (3) numerical treatment of complex propagation media models (e.g. physical dispersion models), (4) algorithmic adaptation to modern high performance computing platforms.

3.2. High order discretization methods

3.2.1. The Discontinuous Galerkin method

The Discontinuous Galerkin method (DG) was introduced in 1973 by Reed and Hill to solve the neutron transport equation. From this time to the 90's a review on the DG methods would likely fit into one page. In the meantime, the Finite Volume approach (FV) has been widely adopted by computational fluid dynamics scientists and has now nearly supplanted classical finite difference and finite element methods in solving problems of nonlinear convection and conservation law systems. The success of the FV method is due to its ability to capture discontinuous solutions which may occur when solving nonlinear equations or more simply, when convecting discontinuous initial data in the linear case. Let us first remark that DG methods share with FV methods this property since a first order FV scheme may be viewed as a 0th order DG scheme. However a DG method may also be considered as a Finite Element (FE) one where the continuity constraint at an element interface is released. While keeping almost all the advantages of the FE method (large spectrum of applications, complex geometries, etc.), the DG method has other nice properties which explain the renewed interest it gains in various domains in scientific computing as witnessed by books or special issues of journals dedicated to this method [42]- [43]- [44]- [49]:

- It is naturally adapted to a high order approximation of the unknown field. Moreover, one may increase the degree of the approximation in the whole mesh as easily as for spectral methods but, with a DG method, this can also be done very locally. In most cases, the approximation relies on a polynomial interpolation method but the DG method also offers the flexibility of applying local approximation strategies that best fit to the intrinsic features of the modeled physical phenomena.
- When the space discretization is coupled to an explicit time integration scheme, the DG method leads to a block diagonal mass matrix whatever the form of the local approximation (e.g. the type of polynomial interpolation). This is a striking difference with classical, continuous FE formulations. Moreover, the mass matrix may be diagonal if the basis functions are orthogonal.
- It easily handles complex meshes. The grid may be a classical conforming FE mesh, a non-conforming one or even a hybrid mesh made of various elements (tetrahedra, prisms, hexahedra, etc.). The DG method has been proven to work well with highly locally refined meshes. This property makes the DG method more suitable (and flexible) to the design of some hp-adaptive solution strategy.
- It is also flexible with regards to the choice of the time stepping scheme. One may combine the DG spatial discretization with any global or local explicit time integration scheme, or even implicit, provided the resulting scheme is stable.
- It is naturally adapted to parallel computing. As long as an explicit time integration scheme is used, the DG method is easily parallelized. Moreover, the compact nature of DG discretization schemes is in favor of high computation to communication ratio especially when the interpolation order is increased.

As with standard FE methods, a DG method relies on a variational formulation of the continuous problem at hand. However, due to the discontinuity of the global approximation, this variational formulation has to be defined locally, at the element level. Then, a degree of freedom in the design of a DG method stems from the approximation of the boundary integral term resulting from the application of an integration by parts to the element-wise variational form. In the spirit of FV methods, the approximation of this boundary integral term calls for a numerical flux function which can be based on either a centered scheme or an upwind scheme, or a blending between these two schemes.

3.2.2. High order DG methods for wave propagation models

DG methods are et the heart of the activities of the pteam regarding the development of high order discretization schemes for the PDE systems modeling electromagnetic and elatsodynamic wave propagation:

- **Nodal DG methods for time-domain problems**. For the numerical solution of the time-domain Maxwell equations, we have first proposed a non-dissipative high order DGTD (Discontinuous Galerkin Time Domain) method working on unstructured conforming simplicial meshes [19]-[2]. This DG method combines a central numerical flux function for the approximation of the integral term at the interface of two neighboring elements with a second order leap-frog time integration scheme. Moreover, the local approximation of the electromagnetic field relies on a nodal (Lagrange type) polynomial interpolation method. Recent achievements by the team deal with the extension of these methods towards non-conforming meshes and hp-adaptivity [16]-[17], their coupling with hybrid explicit/implicit time integration schemes in order to improve their efficiency in the context of locally refined meshes [6]. A high order DG method has also been proposed for the numerical resolution of the elastodynamic equations modeling the propagation of seismic waves [4]-[15].
- **Hybridizable DG (HDG) method for time-domain and time-harmonic problems**. For the numerical treatment of the time-harmonic Maxwell equations, nodal DG methods can also be consiered [7]-[14]. However, such DG formulations are highly expensive, especially for the discretization of 3d problems, because they lead to a large sparse and undefinite linear system of equations coupling all the degrees of freedom of the unknown physical fields. Different attempts have been made in the recent past to improve this situation and one promising strategy has been recently proposed by Cockburn *et al.*[47] in the form of so-called hybridizable DG formulations. The distinctive feature of these methods is that the only globally coupled degrees of freedom are those of an approximation of the solution defined only on the boundaries of the elements. This work is concerned with the study of such Hybridizable Discontinuous Galerkin (HDG) methods for the solution of the system of Maxwell equations in the time-domain when the time integration relies on an implicit scheme, or in the frequency domain. The team has been a precursor in the development of HDG methods for the frequency-domain Maxwell equations [22]-[23].
- **Multiscale DG methods for time-domain problems**. More recently, in the framework of a collaboration with LNCC in Petropolis (Frédéric Valentin), we have started to investigate a family of methods specifically designed for an accurate and efficient numerical treatment of multiscale wave propagation problems. These methods, referred to as Multiscale Hybrid Mixed (MHM) methods, are currently studied in the team for both time-domain electromagnetic and elastodynamic PDE models. They consist in reformulating the mixed variational form of each system into a global (arbitrarily coarse) problem related to a weak formulation of the boundary condition (carried by a Lagrange multiplier that represents e.g. the normal stress tensor in elastodynamic sytems), and a series of small, element-wise, fully decoupled problems resembling to the initial one and related to some well chosen partition of the solution variables on each element. By construction, that methodology is fully parallelizable and recursivity may be used in each local problem as well, making MHM methods belonging to multi-level highly parallelizable methods. Each local problem may be solved using DG or classical Galerkin FE approximations combined with some appropriate time integration scheme (θ-scheme or leap-frog scheme).

3.3. Efficient time integration strategies

The use of unstructured meshes (based on triangles in two space dimensions and tetrahedra in three space dimensions) is an important feature of the DGTD methods developed in the team which can thus easily deal with complex geometries and heterogeneous propagation media. Moreover, DG discretization methods are naturally adapted to local, conforming as well as non-conforming, refinement of the underlying mesh, Most of the existing DGTD methods rely on explicit time integration schemes and lead to block diagonal mass matrices which is often recognized as one of the main advantages with regards to continuous finite element methods. However, explicit DGTD methods are also constrained by a stability condition that can be very restrictive

on highly refined meshes and when the local approximation relies on high order polynomial interpolation. There are basically three strategies that can be considered to cure this computational efficiency problem. The first approach is to use an unconditionally stable implicit time integration scheme to overcome the restrictive constraint on the time step for locally refined meshes. In a second approach, a local time stepping strategy is combined with an explicit time integration scheme. In the third approach, the time step size restriction is overcome by using a hybrid explicit-implicit procedure. In this case, one blends a time implicit and a time explicit schemes where only the solution variables defined on the smallest elements are treated implicitly. The first and third options are considered in the team in the framework of DG [6]-[25]-[24] and HDG [20] discretization methods.

3.4. Numerical treatment of complex material models

Towards the general aim of being able to consider concrete physical situations, we are interested in taking into account in the numerical methodologies that we study, a better description of the propagation of waves in realistic media. In the case of electromagnetics, a typical physical phenomenon that one has to consider is *dispersion*. It is present in almost all media and traduces the way the material reacts to the presence of electromagnetic waves. In the presence of an electric field a medium does not react instantaneously and thus presents an electric polarization of the molecules or electrons that itself influences the electric displacement. In the case of a linear homogeneous isotropic media, there is a linear relation between the applied electric field and the polarization. However, above some range of frequencies (depending on the considered material), the dispersion phenomenon cannot be neglected and the relation between the polarization and the applied electric field becomes complex. This is traduced by a frequency-dependent complex permittivity. Several such models for the characterization of the permittivity exist. Concerning biological media, the Debye model is commonly adopted in the presence of water, biological tissues and polymers, so that it already covers a wide range of applications [21]. If one is interested in modeling the dispersion effects on metals on the nanometer scale and at optical frequencies, which are the conditions that one has to deal with in the context of nanoplasmonics, then the Drude or the Drude-Lorentz models are generally adopted [26]. In the context of seismic wave propagation, we are interested by the intrinsic attenuation of the medium. In realistic configurations, for instance in sedimentary basins where the waves are trapped, we can observe site effects due to local geological and geotechnical conditions which result in a strong increase in amplification and duration of the ground motion at some particular locations. During the wave propagation in such media, a part of the seismic energy is dissipated because of anelastic losses relied to the internal friction of the medium. For these reasons, numerical simulations based on the basic assumption of linear elasticity are no more valid since this assumption result in a severe overestimation of amplitude and duration of the ground motion, even when we are not in presence of a site effect, since intrinsic attenuation is not taken into account.

3.5. High performance numerical computing

Beside basic research activities related to the design of numerical methods and resolution algorithms for the wave propagation models at hand, the team is also committed to demonstrate the benefits of the proposed numerical methodologies in the simulation of challenging three-dimensional problems pertaining to computational electromagnetics and computation geoseismics. For such applications, parallel computing is a mandatory path. Nowadays, modern parallel computers most often take the form of clusters of heterogeneous multiprocessor systems, combining multiple core CPUs with accelerator cards (e.g Graphical Processing Units - GPUs), with complex hierarchical distributed-shared memory systems. Developing numerical algorithms that efficiently exploit such high performance computing architectures raises several challenges, especially in the context of a massive parallelism. In this context, current efforts of the team are towards the exploitation of multiple levels of parallelism (computing systems combining CPUs and GPUs) through the study of hierarchical SPMD (Single Program Multiple Data) strategies for the parallelization of unstructured mesh based solvers.

NANO-D Project-Team (section vide)

NECS Project-Team

3. Research Program

3.1. Introduction

NECS team deals with Networked Control Systems. Since its foundation in 2007, the team has been addressing issues of control under imperfections and constraints deriving from the network (limited computation resources of the embedded systems, delays and errors due to communication, limited energy resources), proposing co-design strategies. The team has recently moved its focus towards general problems on *control of network systems*, which involve the analysis and control of dynamical systems with a network structure or whose operation is supported by networks. This is a research domain with substantial growth and is now recognized as a priority sector by the IEEE Control Systems Society: IEEE has started in a new journal, IEEE Transactions on Control of Network Systems, whose first issue appeared in 2014.

More in detail, the research program of NECS team is along lines described in the following sections.

3.2. Distributed estimation and data fusion in network systems

This research topic concerns distributed data combination from multiple sources (sensors) and related information fusion, to achieve more specific inference than could be achieved by using a single source (sensor). It plays an essential role in many networked applications, such as communication, networked control, monitoring, and surveillance. Distributed estimation has already been considered in the team. We wish to capitalize and strengthen these activities by focusing on integration of heterogeneous, multidimensional, and large data sets:

- Heterogeneity and large data sets. This issue constitutes a clearly identified challenge for the future. Indeed, heterogeneity comes from the fact that data are given in many forms, refer to different scales, and carry different information. Therefore, data fusion and integration will be achieved by developing new multi-perception mathematical models that can allow tracking continuous (macroscopic) and discrete (microscopic) dynamics under a unified framework while making different scales interact with each other. More precisely, many scales are considered at the same time, and they evolve following a unique fully-integrated dynamics generated by the interactions of the scales. The new multi-perception models will be integrated to forecast, estimate and broadcast useful system states in a distributed way. Targeted applications include traffic networks and navigation, and concern recent grant proposals that team has elaborated, among which the SPEEDD EU FP7 project, which has been accepted and started in February 2014 and the LOCATE-ME project, which treats pedestrian navigation.
- Multidimensionality. This issue concerns the analysis and the processing of multidimensional data, organized in multiway array, in a distributed way. Robustness of previously-developed algorithms will be studied. In particular, the issue of missing data will be taken into account. In addition, since the considered multidimensional data are generated by dynamic systems, dynamic analysis of multiway array (or tensors) will be considered. The targeted applications concern distributed detection in complex networks and distributed signal processing for collaborative networks. This topic is developed in strong collaboration with UFC (Brazil).

3.3. Networked systems and graph analysis

This is a research topic at the boundaries between graph theory and dynamical systems theory.

A first main line of research will be to study complex systems whose interactions are modeled with graphs, and to unveil the effect of the graph topology on system-theoretic properties such as observability or controllability. In particular, on-going work concerns observability of graph-based systems: after preliminary results concerning consensus systems over distance-regular graphs, the aim is to extend results to more general networks. A special focus will be on the notion of 'generic properties', namely properties which depend only on the underlying graph describing the sparsity pattern, and hold true almost surely with a random choice of the non-zero coefficients. Further work will be to explore situations in which there is the need for new notions different from the classical observability or controllability. For example, in social networks or in birds flocking the potential leader might have a goal different from classical controllability, because on the one hand he might have a goal much less ambitious than being able to drive the system to any possible state (e.g., he might want to drive everybody near its own opinion, only), and on the other hand he might have much weaker tools to construct its input (e.g., he might not know the whole system's dynamics, but only a few things, possibly that the system is linear and one row of the matrix only). Another example is the question of detectability of an unknown input under the assumption that such an input has a sparsity constraint, a question arising from the fact that a cyber-physical attack might be modeled as an input aiming at controlling the system's state, and that limitations in the capabilities of the attacker might be modeled as a sparsity constraint on the input.

A second line of research will concern graph discovery, namely algorithms aiming at reconstructing some properties of the graph (such as the number of vertices, the diameter, the degree distribution, or spectral properties such as the eigenvalues of the graph Laplacian), using some measurements of quantities related to a dynamical system associated with the graph. It will be particularly challenging to consider directed graphs, and to impose that the algorithm is anonymous, i.e., that it does not makes use of labels identifying the different agents associated with vertices.

3.4. Collaborative and distributed network control

This research line deals with the problem of designing controllers with a limited use of the network information (i.e. with restricted feedback), and with the aim to reach a pre-specified global behavior. This is in contrast to centralized controllers that use the whole system information and compute the control law at some central node. Collaborative control has already been explored in the team in connection with the underwater robot fleet, and to some extent with the source seeking problem. It remains however a certain number of challenging problems that the team wishes to address:

- Design of control with limited information, able to lead to desired global behaviors. Here the graph structure is imposed by the problem, and we aim to design the "best" possible control under such a graph constraint [0]. The team would like to explore further this research line, targeting a better understanding of possible metrics to be used as a target for optimal control design. In particular, and in connection with the traffic application, the long-standing open problem of ramp metering control under minimum information will be addressed.
- Clustering control for large networks. For large and complex systems composed of several sub-networks, feedback design is usually treated at the sub-network level, and most of the times without taking into account natural interconnections between sub-networks. The team wishes to explore new control strategies, exploiting the emergent behaviors resulting from new interconnections between the network components. This requires first to build network models operating in aggregated clusters, and then to re-formulate problems where the control can be designed using the cluster boundaries rather than individual control loops inside of each network. Examples can be found in the transportation application domain, where a significant challenge will be to obtain dynamic partitioning and clustering of heterogeneous networks in homogeneous sub-networks, and then to control the perimeter flows of the clusters to optimize the network operation.

[0] Such a problem has been previously addressed in some specific applications, particularly robot fleets, and only few recent theoretical works have initiated a more systematic system-theoretic study of sparsity-constrained system realization theory and of sparsity-constrained feedback control

3.5. Transportation networks

This is currently the main application domain of the NECS team. Several interesting problems in this area capture many of the generic networks problems described above. For example, distributed collaborative algorithms can be devised for ramp-metering control and traffic-density balancing can be achieved using consensus concepts. The team is already strongly involved in this field, both this theoretical works on traffic prediction and control, and with the Grenoble Traffic Lab platform. These activities will be continued and strengthened.

NON-A Project-Team

3. Research Program

3.1. General annihilators

Estimation is quite easy in the absence of perturbations. It becomes challenging in more realistic situations, faced to measurement noises or other unknown inputs. In our works, as well as in the founding text of *Non-A*, we have shown how our estimation techniques can successfully get rid of perturbations of the so-called *structured* type, which means the ones that can be annihilated by some linear differential operator (called the annihilator). *ALIEN* already defined such operators by integral operators, but using more general convolution operators is an alternative to be analyzed, as well as defining the "best way to kill" perturbations. Open questions are:

OQ1) Does a normal form exist for such annihilators?

OQ2) Or, at least, does there exist an adequate basis representation of the annihilator in some adequate algebra?

OQ3) And lastly, can the annihilator parameters be derived from efficient tuning rules?

The two first questions will directly impact Indicators 1 (time) and 2 (complexity), whereas the last one will impact indicator 3 (robustness).

3.2. Numerical differentiation

Estimating the derivative of a (noisy) signal with a sufficient accuracy can be seen as a key problem in domains of control and diagnosis, as well as signal and image processing. At the present stage of our research, the estimation of the n-th order time derivatives of noisy signals (including noise filtering for $n = 0$) appears as a common area for the whole project, either as a research field, or as a tool that is used both for model-based and model-free techniques. *One of the open questions is about the robustness issues (Indicator 3) with respect to the annihilator, the parameters and the numerical implementation choices.*

Two classes of techniques are considered here (**Model-based** and **Model-free**), both of them aiming at non-asymptotic estimation.

In what we call *model-based techniques*, the derivative estimation is regarded as an observation problem, which means the software-based reconstruction of unmeasured variables and, more generally, a left inversion problem [0]. This involves linear/homogeneous/nonlinear state models, including ordinary equations, systems with delays, hybrid systems with impulses or switches [0], which still has to be exploited in the finite-time and fixed-time context. Power electronics is already one of the possible applications.

Model-free techniques concern the works initiated by *ALIEN*, which rely on the only information contained in the output signal and its derivatives. The corresponding algorithms rely on our algebraic annihilation viewpoint. *One open question is: How to provide an objective comparison analysis between Model-based and Model-free estimation techniques? For this, we will only concentrate on Non-Asymptotic ones. This comparison will have to be based on the three Indicators 1 (time), 2 (complexity) and 3 (robustness).*

[0]Left invertibility deals with the question of recovering the full state of a system ("observation") together with some of its inputs ("unknown input observers"), and also refers to algebraic structural conditions.

[0]Note that hybrid dynamical systems (HDS) constitute an important field of investigation since, in this case, the discrete state can be considered as an unknown input.

3.3. Model-free control

Industry is keen on simple and powerful controllers: the tuning simplicity of the classical PID controller explains its omnipresence in industrial control systems, although its performances drop when working conditions change. The last challenge we consider is to define control techniques which, instead of using sophisticated models (the development of which may be expensive), use the information contained in the output signal and its estimated derivatives, which can be regarded as "signal-based" controllers. *Such design should take into account the Indicators 1 (time), 2 (complexity) and 3 (robustness).*

3.4. Applications

Keeping in mind that we will remain focused at developing and applying fundamental methods for non-asymptotic estimation, we intend to deal with 4 main domains of application (see the lower part of Figure 1). The Lille context offers interesting opportunities in WSAN (wireless sensor and actuator networks and, more particularly, networked robots) at Inria, as well as nano/macro machining at ENSAM. A power electronics platform will be developed in ENSEA Cergy. Last, in contact with companies, several grants, patents and collaborations are expected from the applications of $i-$PID. Each of these four application domains was presented in the *Non-A* proposal:

- Networked robots, WSAN [Lille]
- Nano/macro machining [Lille]
- Multicell chopper [Lille and Cergy]
- *i*-PID for industry

In the present period, we choose to give a particular focus to the first item (Networked robots), which already received some development. It can be considered as the objective 4.

These applications are described with more details below.

OPALE Project-Team

3. Research Program

3.1. Functional and numerical analysis of PDE systems

Our common scientific background is the functional and numerical analysis of PDE systems, in particular with respect to nonlinear hyperbolic equations such as conservation laws of gas-dynamics.

Whereas the structure of weak solutions of the Euler equations has been thoroughly discussed in both the mathematical and fluid mechanics literature, in similar hyperbolic models, focus of new interest, such as those related to traffic, the situation is not so well established, except in one space dimension, and scalar equations. Thus, the study of such equations is one theme of emphasis of our research.

The well-developed domain of numerical methods for PDE systems, in particular finite volumes, constitute the sound background for PDE-constrained optimization.

3.2. Numerical optimization of PDE systems

Partial Differential Equations (PDEs), finite volumes/elements, geometrical optimization, optimum shape design, multi-point/multi-criterion/multi-disciplinary optimization, shape parameterization, gradient-based/evolutionary/hybrid optimizers, hierarchical physical/numerical models, Proper Orthogonal Decomposition (POD)

Optimization problems involving systems governed by PDEs, such as optimum shape design in aerodynamics or electromagnetics, are more and more complex in the industrial setting.

In certain situations, the major difficulty resides in the costly evaluation of a functional by means of a simulation, and the numerical method to be used must exploit at best the problem characteristics (regularity or smoothness, local convexity).

In many other cases, several criteria are to be optimized and some are non differentiable and/or non convex. A large set of parameters, sometimes of different types (boolean, integer, real or functional), are to be taken into account, as well as constraints of various types (physical and geometrical, in particular). Additionally, today's most interesting optimization pre-industrial projects are multi-disciplinary, and this complicates the mathematical, physical and numerical settings. Developing *robust optimizers* is therefore an essential objective to make progress in this area of scientific computing.

In the area of numerical optimization algorithms, the project aims at adapting classical optimization methods (simplex, gradient, quasi-Newton) when applicable to relevant engineering applications, as well as developing and testing less conventional approaches such as Evolutionary Strategies (ES), including Genetic or Particle-Swarm Algorithms, or hybrid schemes, in contexts where robustness is a very severe constraint.

In a different perspective, the heritage from the former project Sinus in Finite-Volumes (or -Elements) for nonlinear hyperbolic problems, leads us to examine cost-efficiency issues of large shape-optimization applications with an emphasis on the PDE approximation; of particular interest to us:

- best approximation and shape-parameterization,
- convergence acceleration (in particular by multi-level methods),
- model reduction (e.g. by *Proper Orthogonal Decomposition*),
- parallel and grid computing; etc.

3.3. Geometrical optimization

Jean-Paul Zolesio and Michel Delfour have developed, in particular in their book [6], a theoretical framework for for geometrical optimization and shape control in Sobolev spaces.

In preparation to the construction of sound numerical techniques, their contribution remains a fundamental building block for the functional analysis of shape optimization formulations.

3.4. Integration platforms

Developing grid, cloud and high-performance computing for complex applications is one of the priorities of the IST chapter in the 7th Framework Program of the European Community. One of the challenges of the 21st century in the computer science area lies in the integration of various expertise in complex application areas such as simulation and optimization in aeronautics, automotive and nuclear simulation. Indeed, the design of the reentry vehicle of a space shuttle calls for aerothermal, aerostructure and aerodynamics disciplines which all interact in hypersonic regime, together with electromagnetics. Further, efficient, reliable, and safe design of aircraft involve thermal flows analysis, consumption optimization, noise reduction for environmental safety, using for example aeroacoustics expertise.

The integration of such various disciplines requires powerful computing infrastructures and particular software coupling techniques. Simultaneously, advances in computer technology militate in favor of the use of massively parallel clusters including hundreds of thousands of processors connected by high-speed gigabits/sec networks. This conjunction makes it possible for an unprecedented cross-fertilization of computational methods and computer science. New approaches including evolutionary algorithms, parameterization, multi-hierarchical decomposition lend themselves seamlessly to parallel implementations in such computing infrastructures. This opportunity is being dealt with by the Opale project-team since its very beginning. A software integration platform has been designed by the Opale project-team for the definition, configuration and deployment of multidisciplinary applications on a distributed heterogeneous infrastructure. Experiments conducted within European projects and industrial cooperations using CAST have led to significant performance results in complex aerodynamics optimization test-cases involving multi-elements airfoils and evolutionary algorithms, i.e. coupling genetic and hierarchical algorithms involving game strategies [77].

The main difficulty still remains however in the deployment and control of complex distributed applications by the end-users. Indeed, the deployment of the computing infrastructures and of the applications in such environments still requires specific expertise by computer science specialists. However, the users, which are experts in their particular application fields, e.g. aerodynamics, are not necessarily experts in distributed and grid computing. Being accustomed to Internet browsers, they want similar interfaces to interact with high-performance computing and problem-solving environments. A first approach to solve this problem is to define component-based infrastructures, e.g. the Corba Component Model, where the applications are considered as connection networks including various application codes. The advantage is here to implement a uniform approach for both the underlying infrastructure and the application modules. However, it still requires specific expertise not directly related to the application domains of each particular user. A second approach is to make use of web services, defined as application and support procedures to standardize access and invocation to remote support and application codes. This is usually considered as an extension of Web services to distributed infrastructures. A new approach, which is currently being explored by the Opale project, is the design of a virtual computing environment able to hide the underlying high-performance-computing infrastructures to the users. The team is exploring the use of distributed workflows to define, monitor and control the execution of high-performance simulations on distributed clusters. The platform includes resilience, i.e., fault-tolerance features allowing for resource demanding and erroneous applications to be dynamically restarted safely, without user intervention.

POEMS Project-Team

3. Research Program

3.1. General description

Our activity relies on the existence of boundary value problems established by physicists to model the propagation of waves in various situations. The basic ingredient is a linear partial differential equation of the hyperbolic type, whose prototype is the wave equation (or the Helmholtz equation if time-periodic solutions are considered). Nowadays, the numerical techniques for solving the basic academic problems are well mastered. However, the resolution of complex wave propagation problems close to real applications still poses (essentially open) problems which constitute a real challenge for applied mathematicians. In particular, several difficulties arise when extending the results and the methods from the scalar wave equation to vectorial problems modeling wave propagation in electromagnetism or elastodynamics.

A large part of research in mathematics, when applied to wave propagation problems, is oriented towards the following goals:

- The conception of new numerical methods, more and more accurate and high performing.
- The development of artificial transparent boundary conditions for handling unbounded propagation domains.
- The treatment of more and more complex problems (non local models, non linear models, coupled systems, periodic media).
- The study of specific phenomena such as guided waves and resonances, which raise mathematical questions of spectral theory.
- The development of approximate models via asymptotic analysis with multiple scales (thin layers, boundary or interfaces, small homogeneities, homogenization, ...).
- The development and the analysis of algorithms for inverse problems (in particular for inverse scattering problems) and imaging techniques, using wave phenomena.

3.2. Wave propagation in non classical media

Extraordinary phenomena regarding the propagation of electromagnetic or acoustic waves appear in materials which have non classical properties: materials with a complex periodic microstructure which behave as a material with negative physical parameters, metals which have a negative dielectric permittivity at optical frequencies, magnetized plasmas which present a strongly anisotropic permittivity tensor with eigenvalues of different signs. These non classical materials raise original questions from theoretical and numerical points of view.

The objective is to study the well-posedness in this unusual context where physical parameters are sign-changing. New functional frameworks must be introduced, due, for instance, to hypersingularities of the electromagnetic field which appear at corners of metamaterials. This has of course numerical counterparts. In particular, classical Perfectly Matched Layers are unstable in these dispersive media, and new approaches must be developed.

Two ANR projects (METAMATH and CHROME) are related to this activity.

3.3. Wave propagation in heterogeneous media

One objective is to develop efficient numerical approaches for the propagation of waves in heterogeneous media.

We aim on one hand to improve homogenized modeling of periodic media, by deriving enriched boundary conditions (or transmission conditions if the periodic structure is embedded in a homogeneous matrix) which take into account the boundary layers phenomena.

On the other hand, we like to develop multi-scale numerical methods when the assumption of periodicity on the spatial distribution of the heterogeneities is either relaxed, or even completely lost. The general idea consists in a coupling between a macroscopic solver, based on a coarse mesh, with some microscopic representation of the field. This latter can be obtained by a numerical microscopic solver or by an analytical asymptotic expansion. This leads to two very different approaches which may be relevant for very different applications.

3.4. Spectral theory and modal approaches for waveguides

The study of waveguides is an old and major topic of the team. Concerning the selfadjoint spectral theory for open waveguides, we turned recently to the very important case of periodic media. One objective is to design periodic structures with localized perturbations to create gaps in the spectrum, containing isolating eigenvalues.

Then, we would like to go further in proving the absence of localized modes in non uniform open waveguides. An original approach has been successfully applied to the scalar problem of a 2D junction. The challenge now is to extend these ideas to other configurations: 3D junctions, bent waveguides, vectorial problems...

Besides, we will continue our activity on modal methods for closed waveguides. In particular, we aim at extending the enriched modal method to take into account curvature and rough boundaries.

Finally, we are developing asymptotic models for networks of thin waveguides which arise in several applications (electric networks, simulation of lung, nanophotonics...).

3.5. Inverse problems

Building on the strong expertise of POEMS in the mathematical modeling of waves, most of our contributions aim at improving inverse scattering methodologies.

We acquired some expertise on the so called Linear Sampling Method, from both the theoretical and the practical points of view. Besides, we are working on topological derivative methods, which exploit small-defect asymptotics of misfit functionals and can thus be viewed as an alternative sampling approach, which can take benefit of our expertise on asymptotic methods.

An originality of our activity is to consider inverse scattering in waveguides (the inverse scattering community generally considers only free-space configurations). This is motivated at the same time by specific issues concerning the ill-posedness of the identification process and by applications to non-destructive techniques, for waveguide configurations (cables, pipes, plates etc...).

Lastly, we continue our work on the so-called exterior approach for solving inverse obstacle problems, which associates quasi-reversibility and level set methods. The objective is now to extend it to evolution problems.

3.6. Integral equations

Our activity in this field aims at developing accurate and fast methods for 3D problems.

On one hand, we developed a systematic approach to the analytical evaluation of singular integrals, which arise in the computation of the matrices of integral equations when two elements of the mesh are either touching each other or geometrically close.

On the other hand, POEMS is developing a Fast Multipole Boundary Element Method (FM-BEM) for elastodynamics, for applications to soil-structure interaction or seismology.

Finally, a posteriori error analysis methodologies and adaptivity for boundary integral equation formulations of acoustic, electromagnetic and elastic wave propagation is investigated in the framework of the ANR project RAFFINE.

3.7. Domain decomposition methods

This is a come back to a topic in which POEMS contributed in the 1990's. It is motivated by our collaborations with the CEA-CESTA and the CEA-LIST, for the solution of large problems in time-harmonic electromagnetism and elastodynamics.

We combine in an original manner classical ideas of Domain Decomposition Methods with the specific formulations that we use for wave problems in unbounded domains, taking benefit of the available analytical representations of the solution (integral representation, modal expansion etc...).

QUANTIC Team

3. Research Program

3.1. Hardware-efficient quantum information processing

The research activities of this section and those in next sections are done in collaboration with the permanent researchers of the future QUANTIC project-team, members of Laboratoire Pierre Aigrain, Benjamin Huard (CNRS) and François Mallet (UPMC), and of Centre Automatique et Systèmes, Pierre Rouchon (Mines Paristech). They have benefited from important scientific exchanges and collaborations with the teams of Serge Haroche, Jean-Michel Raimond and Michel Brune at Laboratoire Kastler Brossel (LKB) and Collège de France and those of Michel Devoret and Robert Schoelkopf at the department of Applied Physics of Yale University.

In this scientific program, we will explore various theoretical and experimental issues concerning protection and manipulation of quantum information. Indeed, the next, critical stage in the development of Quantum Information Processing (QIP) is most certainly the active quantum error correction (QEC). Through this stage one designs, possibly using many physical qubits, an encoded logical qubit which is protected against major decoherence channels and hence admits a significantly longer effective coherence time than a physical qubit. Reliable (fault-tolerant) computation with protected logical qubits usually comes at the expense of a significant overhead in the hardware (up to thousands of physical qubits per logical qubit). Each of the involved physical qubits still needs to satisfy the best achievable properties (coherence times, coupling strengths and tunability). More remarkably, one needs to avoid undesired interactions between various subsystems. This is going to be a major difficulty for qubits on a single chip.

The usual approach for the realization of QEC is to use many qubits to obtain a larger Hilbert space of the qubit register [72], [75]. By redundantly encoding quantum information in this Hilbert space of larger dimension one make the QEC tractable: different error channels lead to distinguishable error syndromes. There are two major drawbacks in using multi-qubit registers. The first, fundamental, drawback is that with each added physical qubit, several new decoherence channels are added. Because of the exponential increase of the Hilbert's space dimension versus the linear increase in the number of decay channels, using enough qubits, one is able to eventually protect quantum information against decoherence. However, multiplying the number of possible errors, this requires measuring more error syndromes. Note furthermore that, in general, some of these new decoherence channels can lead to correlated action on many qubits and this needs to be taken into account with extra care: in particular, such kind of non-local error channels are problematic for surface codes. The second, more practical, drawback is that it is still extremely challenging to build a register of more than on the order of 10 qubits where each of the qubits is required to satisfy near the best achieved properties: these properties include the coherence time, the coupling strengths and the tunability. Indeed, building such a register is not merely only a fabrication task but rather, one requirers to look for architectures such that, each individual qubit can be addressed and controlled independently from the others. One is also required to make sure that all the noise channels are well-controlled and uncorrelated for the QEC to be effective.

We have recently introduced a new paradigm for encoding and protecting quantum information in a quantum harmonic oscillator (e.g. a high-Q mode of a 3D superconducting cavity) instead of a multi-qubit register [4]. The infinite dimensional Hilbert space of such a system can be used to redundantly encode quantum information. The power of this idea lies in the fact that the dominant decoherence channel in a cavity is photon damping, and no more decay channels are added if we increase the number of photons we insert in the cavity. Hence, only a single error syndrome needs to be measured to identify if an error has occurred or not. Indeed, we are convinced that most early proposals on continuous variable QIP [49], [45] could be revisited taking into account the design flexibilities of Quantum Superconducting Circuits (QSC) and the new coupling regimes that are provided by these systems. In particular, we have illustrated that coupling a qubit to the cavity mode in the strong dispersive regime provides an important controllability over the Hilbert space of

the cavity mode [51]. Through a recent experimental work [10], we benefit from this controllability to prepare superpositions of quasi-orthogonal coherent states, also known as Schrödinger cat states.

In this Scheme, the logical qubit is encoded in a four-component Schrödinger cat state. Continuous quantum non-demolition (QND) monitoring of a single physical observable, consisting of photon number parity, enables then the tractability of single photon jumps. We obtain therefore a first-order quantum error correcting code using only a single high-Q cavity mode (for the storage of quantum information), a single qubit (providing the non-linearity needed for controllability) and a single low-Q cavity mode (for reading out the error syndrome). As shown in Figure 1 , this leads to a significant hardware economy for realization of a protected logical qubit. Our goal here is to push these ideas towards a reliable and hardware-efficient paradigm for universal quantum computation.

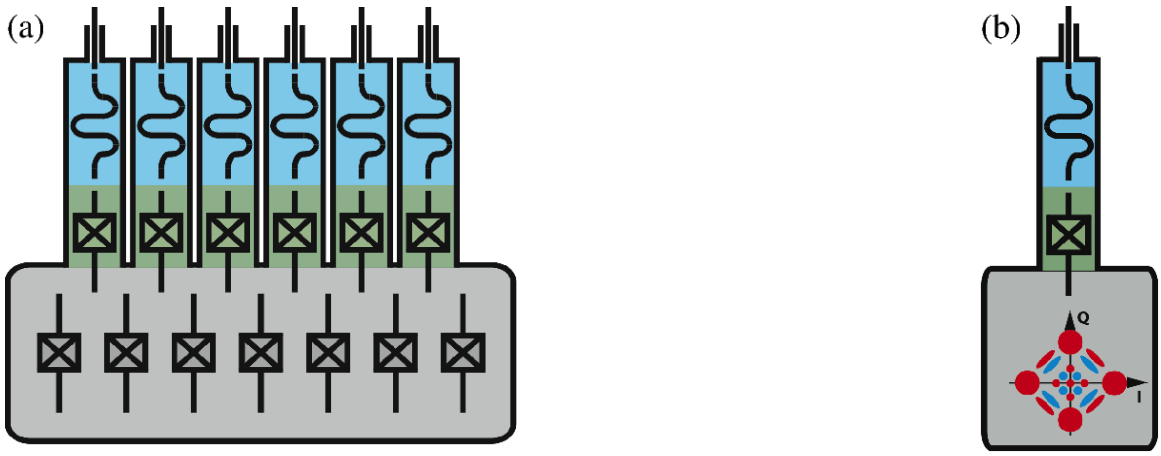


Figure 1. (a) A protected logical qubit consisting of a register of many qubits: here, we see a possible architecture for the Steane code [75] consisting of 7 qubits requiring the measurement of 6 error syndromes. In this sketch, 7 transmon qubits in a high-Q resonator and the measurement of the 6 error syndromes is ensured through 6 additional ancillary qubits with the possibility of individual readout of the ancillary qubits via independent low-Q resonators. (b) Minimal architecture for a protected logical qubit, adapted to circuit quantum electrodynamics experiments. Quantum information is encoded in a Schrödinger cat state of a single high-Q resonator mode and a single error syndrome is measured, using a single ancillary transmpn qubit and the associated readout low-Q resonator.

3.2. Reservoir (dissipation) engineering and autonomous stabilization of quantum systems

Being at the heart of any QEC protocol, the concept of feedback is central for the protection of the quantum information enabling many-qubit quantum computation or long-distance quantum communication. However, such a closed-loop control which requires a real-time and continuous measurement of the quantum system has been for long considered as counter-intuitive or even impossible. This thought was mainly caused by properties of quantum measurements: any measurement implies an instantaneous strong perturbation to the system's state. The concept of *quantum non-demolotion* (QND) measurement has played a crucial role in understanding and resolving this difficulty [30]. In the context of cavity quantum electro-dynamics (cavity QED) with Rydberg atoms [47], a first experiment on continuous QND measurements of the number of microwave photons was performed by the group at Laboratoire Kastler-Brossel (ENS) [46]. Later on, this ability of performing continuous measurements allowed the same group to realize the first continuous quantum

feedback protocol stabilizing highly non-classical states of the microwave field in the cavity, the so-called photon number states [7] (this ground-breaking work was mentioned in the Nobel prize attributed to Serge Haroche). The QUANTIC team contributed to the theoretical work behind this experiment [38], [21], [74], [23]. These contributions include the development and optimization of the quantum filters taking into account the quantum measurement back-action and various measurement noises and uncertainties, the development of a feedback law based on control Lyapunov techniques, and the compensation of the feedback delay.

In the context of circuit quantum electrodynamics (circuit QED) [37], recent advances in quantum-limited amplifiers [67], [77] have opened doors to high-fidelity non-demolition measurements and real-time feedback for superconducting qubits [2]. This ability to perform high-fidelity non-demolition measurements of a quantum signal has very recently led to quantum feedback experiments with quantum superconducting circuits [77], [66], [32]. Here again, the QUANTIC team has participated to one of the first experiments in the field where the control objective is to track a dynamical trajectory of a single qubit rather than stabilizing a stationary state (this experiment was performed by the members of the future QUANTIC project-team). Such quantum trajectory tracking could be further explored to achieve metrological goals such as the stabilization of the amplitude of a microwave drive [58].

While all this progress has led to a strong optimism about the possibility to perform active protection of quantum information against decoherence, the rather short dynamical time scales of these systems limit, to a great amount, the complexity of the feedback strategies that could be employed. Indeed, in such measurement-based feedback protocols, the time-consuming data acquisition and post-treatment of the output signal leads to an important latency in the feedback procedure.

The reservoir (dissipation) engineering [64] and the closely related coherent feedback [56] are considered as alternative approaches circumventing the necessity of a real-time data acquisition, signal processing and feedback calculations. In the context of quantum information, the decoherence, caused by the coupling of a system to uncontrolled external degrees of freedom, is generally considered as the main obstacle to synthesize quantum states and to observe quantum effects. Paradoxically, it is possible to intentionally engineer a particular coupling to a reservoir in the aim of maintaining the coherence of some particular quantum states. In a general viewpoint, these approaches could be understood in the following manner: by coupling the quantum system to be stabilized to a strongly dissipative ancillary quantum system, one evacuates the entropy of the main system through the dissipation of the ancillary one. By building the feedback loop into the Hamiltonian, this type of autonomous feedback obviates the need for a complicated external control loop to correct errors. On the experimental side, such autonomous feedback techniques have been used for qubit reset [1], single-qubit state stabilization [59], and the creation [25] and stabilization [50], [55][8] of states of multipartite quantum systems.

Such reservoir engineering techniques could be widely revisited exploring the flexibility in the Hamiltonian design for QSC. We have recently developed theoretical proposals leading to extremely efficient, and simple to implement, stabilization schemes for systems consisting of a single or two qubits [1] [53]. The experimental results based on these protocols have illustrated the efficiency of the approach [1], [8]. Through these experiments, we exploit the strong dispersive interaction [70] between superconducting qubits and a single low-Q cavity mode playing the role of a dissipative reservoir. Applying some continuous-wave (cw) microwave drives with well-chosen fixed frequencies, amplitudes, and phases, we engineer an effective interaction Hamiltonian which evacuates entropy from the qubits when an eventual perturbation occurs: by driving the qubits and cavity with continuous-wave drives, we induce an autonomous feedback loop which corrects the state of the qubits every time it decays out of the desired target state. The schemes are robust against small variations of the control parameters (drives amplitudes and phase) and require only some basic calibration. Finally, by avoiding resonant interactions between the qubits and the low-Q cavity mode, the qubits remain protected against the Purcell effect, which would reduce the coherence times.

3.3. System theory for quantum information processing

In parallel and in strong interactions with the above experimental goals, we develop systematic mathematical methods for dynamical analysis, control and estimation of composite and open quantum systems. These

systems are built with several quantum subsystems whose irreversible dynamics results from measurements and/or decoherence. A special attention is given to spin/spring systems made with qubits and harmonic oscillators. These developments are done in the spirit of our recent contributions [68], [21], [73], [74], [23][6] [69] resulting from collaborations with the cavity quantum electrodynamics group of Laboratoire Kastler Brossel.

3.3.1. Stabilization by measurement-based feedback

The protection of quantum information via efficient QEC is a combination of (i) tailored dynamics of a quantum system in order to protect an informational qubit from certain decoherence channels, and (ii) controlled reaction to measurements that efficiently detect and correct the dominating disturbances that are not rejected by the tailored quantum dynamics.

In such feedback scheme, the system and its measurement are quantum objects whereas the controller and the control input are classical. The stabilizing control law is based on the past values of the measurement outcomes. During our work on the LKB photon box, we have developed, for single input systems subject to quantum non-demolition measurement, a systematic stabilization method [23]: it is based on a discrete-time formulation of the dynamics, on the construction of a strict control Lyapunov function and on an explicit compensation of the feedback-loop delay. Keeping the QND measurement assumptions, extensions of such stabilization schemes will be investigated in the following directions: finite set of values for the control input with application to the convergence analysis of the atomic feedback scheme experimentally tested in [79]; multi-input case where the construction by inversion of a Metzler matrix of the strict Lyapunov function is not straightforward; continuous-time systems governed by diffusive master equations; stabilization towards a set of density operators included in a target subspace; adaptive measurement by feedback to accelerate the convergence towards a stationary state as experimentally tested in [62]. Without the QND measurement assumptions, we will also address the stabilization of non-stationary states and trajectory tracking, with applications to systems similar to those considered in [2] [32].

3.3.2. Filtering, quantum state and parameter estimations

The performance of every feedback controller crucially depends on its online estimation of the current situation. This becomes even more important for quantum systems, where full state measurements are physically impossible. Therefore the ultimate performance of feedback correction depends on fast, efficient and optimally accurate state and parameter estimations.

A quantum filter takes into account imperfection and decoherence and provides the quantum state at time $t \geq 0$ from an initial value at $t = 0$ and the measurement outcomes between 0 and t. Quantum filtering goes back to the work of Belavkin [26] and is related to quantum trajectories [34], [36]. A modern and mathematical exposure of the diffusive models is given in [24]. In [80] a first convergence analysis of diffusive filters is proposed. Nevertheless the convergence characterization and estimation of convergence rate remain open and difficult problems. For discrete time filters, a general stability result based on fidelity is proven in [68], [73]. This stability result is extended to a large class of continuous-time filters in [22]. Further efforts are required to characterize asymptotic and exponential stability. Estimations of convergence rates are available only for quantum non-demolition measurements [27]. Parameter estimations based on measurement data of quantum trajectories can be formulated within such quantum filtering framework [40], [60].

We will continue to investigate stability and convergence of quantum filtering. We will also exploit our fidelity-based stability result to justify maximum likelihood estimation and to propose, for open quantum system, parameter estimation algorithms inspired of existing estimation algorithms for classical systems. We will also investigate a more specific quantum approach: it is noticed in [31] that post-selection statistics and "past quantum" state analysis [41] enhance sensitivity to parameters and could be interesting towards increasing the precision of an estimation.

3.3.3. Stabilization by interconnections

In such stabilization schemes, the controller is also a quantum object: it is coupled to the system of interest and is subject to decoherence and thus admits an irreversible evolution. These stabilization schemes are closely

related to reservoir engineering and coherent feedback [64], [56]. The closed-loop system is then a composite system built with the original system and its controller. In fact, and given our particular recent expertise in this domain [6], [1], [8], this subsection is dedicated to further developing such stabilization techniques, both experimentally and theoretically.

The main analysis issues are to prove the closed-loop convergence and to estimate the convergence rates. Since these systems are governed by Lindblad differential equations (continuous-time case) or Kraus maps (discrete-time case), their stability is automatically guaranteed: such dynamics are contractions for a large set of metrics (see [63]). Convergence and asymptotic stability is less well understood. In particular most of the convergence results consider the case where the target steady-state is a density operator of maximum rank (see, e.g., [20][chapter 4, section 6]). When the goal steady-state is not full rank very few convergence results are available.

We will focus on this geometric situation where the goal steady-state is on the boundary of the cone of positive Hermitian operators of finite trace. A specific attention will be given to adapt standard tools (Lyapunov function, passivity, contraction and Lasalle's invariance principle) for infinite dimensional systems to spin/spring structures inspired of [6], [1], [8], [5] and their associated Fokker-Planck equations for the Wigner functions.

We will also explore the Heisenberg point of view in connection with recent results of the Inria project-team MAXPLUS (algorithms and applications of algebras of max-plus type) relative to Perron-Frobenius theory [44], [43]. We will start with [71] and [65] where, based on a theorem due to Birkhoff [28], dual Lindblad equations and dual Kraus maps governing the Heisenberg evolution of any operator are shown to be contractions on the cone of Hermitian operators equipped with Hilbert's projective metric. As the Heisenberg picture is characterized by convergence of all operators to a multiple of the identity, it might provide a mean to circumvent the rank issues. We hope that such contraction tools will be especially well adapted to analyzing quantum systems composed of multiple components, motivated by the facts that the same geometry describes the contraction of classical systems undergoing synchronizing interactions [76] and by our recent generalized extension of the latter synchronizing interactions to quantum systems [57].

Besides these analysis tasks, the major challenge in stabilization by interconnections is to provide systematic methods for the design, from typical building blocks, of control systems that stabilize a specific quantum goal (state, set of states, operation) when coupled to the target system. While constructions exist for so-called linear quantum systems [61], this does not cover the states that are more interesting for quantum applications. Various strategies have been proposed that concatenate iterative control steps for open-loop steering [78], [54] with experimental limitations. The characterization of Kraus maps to stabilize any types of states has also been established [29], but without considering experimental implementations. A viable stabilization by interaction has to combine the capabilities of these various approaches, and this is a missing piece that we want to address.

3.3.3.1. Perturbation methods

With this subsection we turn towards more fundamental developments that are necessary in order to address the complexity of quantum networks with efficient reduction techniques. This should yield both efficient mathematical methods, as well as insights towards unravelling dominant physical phenomena/mechanisms in multipartite quantum dynamical systems.

In the Schrödinger point of view, the dynamics of open quantum systems are governed by master equations, either deterministic or stochastic [47], [42]. Dynamical models of composite systems are based on tensor products of Hilbert spaces and operators attached to the constitutive subsystems. Generally, a hierarchy of different timescales is present. Perturbation techniques can be very useful to construct reliable models adapted to the timescale of interest.

To eliminate high frequency oscillations possibly induced by quasi-resonant classical drives, averaging techniques are used (rotating wave approximation). These techniques are well established for closed systems without any dissipation nor irreversible effect due to measurement or decoherence. We will consider in a first step the adaptation of these averaging techniques to deterministic Lindblad master equations governing the quantum state, i.e. the system density operator. Emphasis will be put on first order and higher order corrections

based on non-commutative computations with the different operators appearing in the Lindblad equations. Higher order terms could be of some interest for the protected logical qubit of figure 1 b. In future steps, we intend to explore the possibility to explicitly exploit averaging or singular perturbation properties in the design of coherent quantum feedback systems; this should be an open-systems counterpart of works like [52].

To eliminate subsystems subject to fast convergence induced by decoherence, singular perturbation techniques can be used. They provide reduced models of smaller dimension via the adiabatic elimination of the rapidly converging subsystems. The derivation of the slow dynamics is far from being obvious (see, e.g., the computations of page 142 in [33] for the adiabatic elimination of low-Q cavity). Contrarily to the classical composite systems where we have to eliminate one component in a Cartesian product, we here have to eliminate one component in a tensor product. We will adapt geometric singular perturbations [39] and invariant manifold techniques [35] to such tensor product computations to derive reduced slow approximations of any order. Such adaptations will be very useful in the context of quantum Zeno dynamics to obtain approximations of the slow dynamics on the decoherence-free subspace corresponding to the slow attractive manifold.

Perturbation methods are also precious to analyze convergence rates. Deriving the spectrum attached to the Lindblad differential equation is not obvious. We will focus on the situation where the decoherence terms of the form $L\rho L^\dagger - (L^\dagger L\rho + \rho L^\dagger L)/2$ are small compared to the conservative terms $-i[H/\hbar, \rho]$. The difficulty to overcome here is the degeneracy of the unperturbed spectrum attached to the conservative evolution $\frac{d}{dt}\rho = -i[H/\hbar, \rho]$. The degree of degeneracy of the zero eigenvalue always exceeds the dimension of the Hilbert space. Adaptations of usual perturbation techniques [48] will be investigated. They will provide estimates of convergence rates for slightly open quantum systems. We expect that such estimates will help to understand the dependance on the experimental parameters of the convergence rates observed in [1], [8] [53].

As particular outcomes for the other subsections, we expect that these developments towards simpler dominant dynamics will guide the search for optimal control strategies, both in open-loop microwave networks and in autonomous stabilization schemes such as reservoir engineering. It will further help to efficiently compute explicit convergence rates and quantitative performances for all the intended experiments.

REALOPT Project-Team

3. Research Program

3.1. Introduction

Combinatorial optimization is the field of discrete optimization problems. In many applications, the most important decisions (control variables) are binary (on/off decisions) or integer (indivisible quantities). Extra variables can represent continuous adjustments or amounts. This results in models known as *mixed integer programs* (MIP), where the relationships between variables and input parameters are expressed as linear constraints and the goal is defined as a linear objective function. MIPs are notoriously difficult to solve: good quality estimations of the optimal value (bounds) are required to prune enumeration-based global-optimization algorithms whose complexity is exponential. In the standard approach to solving an MIP is so-called *branch-and-bound algorithm* : (i) one solves the linear programming (LP) relaxation using the simplex method; (ii) if the LP solution is not integer, one adds a disjunctive constraint on a factional component (rounding it up or down) that defines two sub-problems; (iii) one applies this procedure recursively, thus defining a binary enumeration tree that can be pruned by comparing the local LP bound to the best known integer solution. Commercial MIP solvers are essentially based on branch-and-bound (such IBM-CPLEX, FICO-Xpress-mp, or GUROBI). They have made tremendous progress over the last decade (with a speedup by a factor of 60). But extending their capabilities remains a continuous challenge; given the combinatorial explosion inherent to enumerative solution techniques, they remain quickly overwhelmed beyond a certain problem size or complexity.

Progress can be expected from the development of tighter formulations. Central to our field is the characterization of polyhedra defining or approximating the solution set and combinatorial algorithms to identify "efficiently" a minimum cost solution or separate an unfeasible point. With properly chosen formulations, exact optimization tools can be competitive with other methods (such as meta-heuristics) in constructing good approximate solutions within limited computational time, and of course has the important advantage of being able to provide a performance guarantee through the relaxation bounds. Decomposition techniques are implicitly leading to better problem formulation as well, while constraint propagation are tools from artificial intelligence to further improve formulation through intensive preprocessing. A new trend is robust optimization where recent progress have been made: the aim is to produce optimized solutions that remain of good quality even if the problem data has stochastic variations. In all cases, the study of specific models and challenging industrial applications is quite relevant because developments made into a specific context can become generic tools over time and see their way into commercial software.

Our project brings together researchers with expertise in mathematical programming (polyhedral approaches, Dantzig-Wolfe decomposition, mixed integer programing, robust and stochastic programming, and dynamic programming), graph theory (characterization of graph properties, combinatorial algorithms) and constraint programming in the aim of producing better quality formulations and developing new methods to exploit these formulations. These new results are then applied to find high quality solutions for practical combinatorial problems such as routing, network design, planning, scheduling, cutting and packing problems.

3.2. Polyhedral approaches for MIP

Adding valid inequalities to the polyhedral description of an MIP allows one to improve the resulting LP bound and hence to better prune the enumeration tree. In a cutting plane procedure, one attempt to identify valid inequalities that are violated by the LP solution of the current formulation and adds them to the formulation. This can be done at each node of the branch-and-bound tree giving rise to a so-called *branch-and-cut algorithm* [73]. The goal is to reduce the resolution of an integer program to that of a linear program by deriving a linear description of the convex hull of the feasible solutions. Polyhedral theory tells us that if X is a mixed integer program: $X = P \cap \mathbb{Z}^n \times \mathbb{R}^p$ where $P = \{x \in \mathbb{R}^{n+p} : Ax \leq b\}$ with matrix

$(A, b) \in \mathbb{Q}^{m \times (n+p+1)}$, then $conv(X)$ is a polyhedron that can be described in terms of linear constraints, i.e. it writes as $conv(X) = \{x \in \mathbb{R}^{n+p} : C\, x \leq d\}$ for some matrix $(C, d) \in \mathbb{Q}^{m' \times (n+p+1)}$ although the dimension m' is typically quite large. A fundamental result in this field is the equivalence of complexity between solving the combinatorial optimization problem $\min\{cx : x \in X\}$ and solving the *separation problem* over the associated polyhedron $conv(X)$: if $\widetilde{x} \notin conv(X)$, find a linear inequality $\pi\, x \geq \pi_0$ satisfied by all points in $conv(X)$ but violated by $\widetilde{x}$. Hence, for NP-hard problems, one can not hope to get a compact description of $conv(X)$ nor a polynomial time exact separation routine. Polyhedral studies focus on identifying some of the inequalities that are involved in the polyhedral description of $conv(X)$ and derive efficient *separation procedures* (cutting plane generation). Only a subset of the inequalities $C\, x \leq d$ can offer a good approximation, that combined with a branch-and-bound enumeration techniques permits to solve the problem. Using *cutting plane algorithm* at each node of the branch-and-bound tree, gives rise to the algorithm called *branch-and-cut*.

3.3. Decomposition and reformulation approaches

An hierarchical approach to tackle complex combinatorial problems consists in considering separately different substructures (subproblems). If one is able to implement relatively efficient optimization on the substructures, this can be exploited to reformulate the global problem as a selection of specific subproblem solutions that together form a global solution. If the subproblems correspond to subset of constraints in the MIP formulation, this leads to Dantzig-Wolfe decomposition. If it corresponds to isolating a subset of decision variables, this leads to Bender's decomposition. Both lead to extended formulations of the problem with either a huge number of variables or constraints. Dantzig-Wolfe approach requires specific algorithmic approaches to generate subproblem solutions and associated global decision variables dynamically in the course of the optimization. This procedure is known as *column generation*, while its combination with branch-and-bound enumeration is called *branch-and-price*. Alternatively, in Bender's approach, when dealing with exponentially many constraints in the reformulation, the *cutting plane procedures* that we defined in the previous section are well-suited tools. When optimization on a substructure is (relatively) easy, there often exists a tight reformulation of this substructure typically in an extended variable space. This gives rise powerful reformulation of the global problem, although it might be impractical given its size (typically pseudo-polynomial). It can be possible to project (part of) the extended formulation in a smaller dimensional space if not the original variable space to bring polyhedral insight (cuts derived through polyhedral studies can often be recovered through such projections).

3.4. Integration of Artificial Intelligence Techniques in Integer Programming

When one deals with combinatorial problems with a large number of integer variables, or tightly constrained problems, mixed integer programming (MIP) alone may not be able to find solutions in a reasonable amount of time. In this case, techniques from artificial intelligence can be used to improve these methods. In particular, we use primal heuristics and constraint programming.

Primal heuristics are useful to find feasible solutions in a small amount of time. We focus on heuristics that are either based on integer programming (rounding, diving, relaxation induced neighborhood search, feasibility pump), or that are used inside our exact methods (heuristics for separation or pricing subproblem, heuristic constraint propagation, ...).

Constraint Programming (CP) focuses on iteratively reducing the variable domains (sets of feasible values) by applying logical and problem-specific operators. The latter propagates on selected variables the restrictions that are implied by the other variable domains through the relations between variables that are defined by the constraints of the problem. Combined with enumeration, it gives rise to exact optimization algorithms. A CP approach is particularly effective for tightly constrained problems, feasibility problems and min-max problems Mixed Integer Programming (MIP), on the other hand, is known to be effective for loosely constrained problems and for problems with an objective function defined as the weighted sum of variables. Many problems belong to the intersection of these two classes. For such problems, it is reasonable to use algorithms that exploit complementary strengths of Constraint Programming and Mixed Integer Programming.

3.5. Robust Optimization

Decision makers are usually facing several sources of uncertainty, such as the variability in time or estimation errors. A simplistic way to handle these uncertainties is to overestimate the unknown parameters. However, this results in over-conservatism and a significant waste in resource consumption. A better approach is to account for the uncertainty directly into the decision aid model by considering mixed integer programs that involve uncertain parameters. Stochastic optimization account for the expected realization of random data and optimize an expected value representing the average situation. Robust optimization on the other hand entails protecting against the worst-case behavior of unknown data. There is an analogy to game theory where one considers an oblivious adversary choosing the realization that harms the solution the most. A full worst case protection against uncertainty is too conservative and induces very high over-cost. Instead, the realization of random data are bound to belong to a restricted feasibility set, the so-called uncertainty set. Stochastic and robust optimization rely on very large scale programs where probabilistic scenarios are enumerated. There is hope of a tractable solution for realistic size problems, provided one develops very efficient ad-hoc algorithms. The techniques for dynamically handling variables and constraints (column-and-row generation and Bender's projection tools) that are at the core of our team methodological work are specially well-suited to this context.

3.6. Polyhedral Combinatorics and Graph Theory

Many fundamental combinatorial optimization problems can be modeled as the search for a specific structure in a graph. For example, ensuring connectivity in a network amounts to building a *tree* that spans all the nodes. Inquiring about its resistance to failure amounts to searching for a minimum cardinality *cut* that partitions the graph. Selecting disjoint pairs of objects is represented by a so-called *matching*. Disjunctive choices can be modeled by edges in a so-called *conflict graph* where one searches for *stable sets* – a set of nodes that are not incident to one another. Polyhedral combinatorics is the study of combinatorial algorithms involving polyhedral considerations. Not only it leads to efficient algorithms, but also, conversely, efficient algorithms often imply polyhedral characterizations and related min-max relations. Developments of polyhedral properties of a fundamental problem will typically provide us with more interesting inequalities well suited for a branch-and-cut algorithm to more general problems. Furthermore, one can use the fundamental problems as new building bricks to decompose the more general problem at hand. For problem that let themselves easily be formulated in a graph setting, the graph theory and in particular graph decomposition theorem might help.

REGULARITY Project-Team

3. Research Program

3.1. Theoretical aspects: probabilistic modeling of irregularity

The modeling of essentially irregular phenomena is an important challenge, with an emphasis on understanding the sources and functions of this irregularity. Probabilistic tools are well-adapted to this task, provided one can design stochastic models for which the regularity can be measured and controlled precisely. Two points deserve special attention:

- first, the study of regularity has to be *local*. Indeed, in most applications, one will want to act on a system based on local temporal or spatial information. For instance, detection of arrhythmias in ECG or of krachs in financial markets should be performed in "real time", or, even better, ahead of time. In this sense, regularity is a *local* indicator of the *local* health of a system.
- Second, although we have used the term "irregularity" in a generic and somewhat vague sense, it seems obvious that, in real-world phenomena, regularity comes in many colors, and a rigorous analysis should distinguish between them. As an example, at least two kinds of irregularities are present in financial logs: the local "roughness" of the records, and the local density and height of jumps. These correspond to two different concepts of regularity (in technical terms, Hölder exponents and local index of stability), and they both contribute a different manner to financial risk.

In view of the above, the *Regularity* team focuses on the design of methods that:

1. define and study precisely various relevant measures of local regularity,
2. allow to build stochastic models versatile enough to mimic the rapid variations of the different kinds of regularities observed in real phenomena,
3. allow to estimate as precisely and rapidly as possible these regularities, so as to alert systems in charge of control.

Our aim is to address the three items above through the design of mathematical tools in the field of probability (and, to a lesser extent, statistics), and to apply these tools to uncertainty management as described in the following section. We note here that we do not intend to address the problem of controlling the phenomena based on regularity, that would naturally constitute an item 4 in the list above. Indeed, while we strongly believe that generic tools may be designed to measure and model regularity, and that these tools may be used to analyze real-world applications, in particular in the field of uncertainty management, it is clear that, when it comes to control, application-specific tools are required, that we do not wish to address.

The research topics of the *Regularity* team can be roughly divided into two strongly interacting axes, corresponding to two complementary ways of studying regularity:

1. developments of tools allowing to characterize, measure and estimate various notions of local regularity, with a particular emphasis on the stochastic frame,
2. definition and fine analysis of stochastic models for which some aspects of local regularity may be prescribed.

These two aspects are detailed in sections 3.2 and 3.3 below.

3.2. Tools for characterizing and measuring regularity

Fractional Dimensions

Although the main focus of our team is on characterizing *local* regularity, on occasions, it is interesting to use a *global* index of regularity. Fractional dimensions provide such an index. In particular, the *regularization dimension*, that was defined in [31], is well adapted to the study stochastic processes, as its definition allows to build robust estimators in an easy way. Since its introduction, regularization dimension has been used by various teams worldwide in many different applications including the characterization of certain stochastic processes, statistical estimation, the study of mammographies or galactograms for breast carcinomas detection, ECG analysis for the study of ventricular arrhythmia, encephalitis diagnosis from EEG, human skin analysis, discrimination between the nature of radioactive contaminations, analysis of porous media textures, well-logs data analysis, agro-alimentary image analysis, road profile analysis, remote sensing, mechanical systems assessment, analysis of video games, ...(see http://regularity.saclay.inria.fr/theory/localregularity/biblioregdim for a list of works using the regularization dimension).

Hölder exponents

The simplest and most popular measures of local regularity are the pointwise and local Hölder exponents. For a stochastic process $\{X(t)\}_{t\in\mathbb{R}}$ whose trajectories are continuous and nowhere differentiable, these are defined, at a point t_0, as the random variables:

$$\alpha_X(t_0,\omega) = \sup\left\{\alpha : \limsup_{\rho\to 0} \sup_{t,u\in B(t_0,\rho)} \frac{|X_t - X_u|}{\rho^\alpha} < \infty\right\}, \tag{60}$$

and

$$\widetilde{\alpha}_X(t_0,\omega) = \sup\left\{\alpha : \limsup_{\rho\to 0} \sup_{t,u\in B(t_0,\rho)} \frac{|X_t - X_u|}{\|t-u\|^\alpha} < \infty\right\}. \tag{61}$$

Although these quantities are in general random, we will omit as is customary the dependency in ω and X and write $\alpha(t_0)$ and $\widetilde{\alpha}(t_0)$ instead of $\alpha_X(t_0,\omega)$ and $\widetilde{\alpha}_X(t_0,\omega)$.

The random functions $t \mapsto \alpha_X(t_0,\omega)$ and $t \mapsto \widetilde{\alpha}_X(t_0,\omega)$ are called respectively the pointwise and local Hölder functions of the process X.

The pointwise Hölder exponent is a very versatile tool, in the sense that the set of pointwise Hölder functions of continuous functions is quite large (it coincides with the set of lower limits of sequences of continuous functions [6]). In this sense, the pointwise exponent is often a more precise tool (*i.e.* it varies in a more rapid way) than the local one, since local Hölder functions are always lower semi-continuous. This is why, in particular, it is the exponent that is used as a basis ingredient in multifractal analysis (see section 3.2). For certain classes of stochastic processes, and most notably Gaussian processes, it has the remarkable property that, at each point, it assumes an almost sure value [18]. SRP, mBm, and processes of this kind (see sections 3.3 and 3.3) rely on the sole use of the pointwise Hölder exponent for prescribing the regularity.

However, α_X obviously does not give a complete description of local regularity, even for continuous processes. It is for instance insensitive to "oscillations", contrarily to the local exponent. A simple example in the deterministic frame is provided by the function $x^\gamma \sin(x^{-\beta})$, where γ, β are positive real numbers. This so-called "chirp function" exhibits two kinds of irregularities: the first one, due to the term x^γ is measured by the pointwise Hölder exponent. Indeed, $\alpha(0) = \gamma$. The second one is due to the wild oscillations around 0, to which α is blind. In contrast, the local Hölder exponent at 0 is equal to $\frac{\gamma}{1+\beta}$, and is thus influenced by the oscillatory behaviour.

Another, related, drawback of the pointwise exponent is that it is not stable under integro-differentiation, which sometimes makes its use complicated in applications. Again, the local exponent provides here a useful complement to α, since $\widetilde{\alpha}$ is stable under integro-differentiation.

Both exponents have proved useful in various applications, ranging from image denoising and segmentation to TCP traffic characterization. Applications require precise estimation of these exponents.

Stochastic 2-microlocal analysis

Neither the pointwise nor the local exponents give a complete characterization of the local regularity, and, although their joint use somewhat improves the situation, it is far from yielding the complete picture.

A fuller description of local regularity is provided by the so-called *2-microlocal analysis*, introduced by J.M. Bony [46]. In this frame, regularity at each point is now specified by two indices, which makes the analysis and estimation tasks more difficult. More precisely, a function f is said to belong to the *2-microlocal space* $C_{x_0}^{s,s'}$, where $s+s'>0, s'<0$, if and only if its $m=[s+s']-$th order derivative exists around x_0, and if there exists $\delta>0$, a polynomial P with degree lower than $[s]-m$, and a constant C, such that

$$\left|\frac{\partial^m f(x)-P(x)}{|x-x_0|^{[s]-m}}-\frac{\partial^m f(y)-P(y)}{|y-x_0|^{[s]-m}}\right|\leq C|x-y|^{s+s'-m}(|x-y|+|x-x_0|)^{-s'-[s]+m}$$

for all x,y such that $0<|x-x_0|<\delta$, $0<|y-x_0|<\delta$. This characterization was obtained in [25], [32]. See [53], [54] for other characterizations and results. These spaces are stable through integro-differentiation, i.e. $f\in C_x^{s,s'}$ if and only if $f'\in C_x^{s-1,s'}$. Knowing to which space f belongs thus allows to predict the evolution of its regularity after derivation, a useful feature if one uses models based on some kind differential equations. A lot of work remains to be done in this area, in order to obtain more general characterizations, to develop robust estimation methods, and to extend the "2-microlocal formalism" : this is a tool allowing to detect which space a function belongs to, from the computation of the Legendre transform of an auxiliary function known as its *2-microlocal spectrum*. This spectrum provide a wealth of information on the local regularity.

In [18], we have laid some foundations for a stochastic version of 2-microlocal analysis. We believe this will provide a fine analysis of the local regularity of random processes in a direction different from the one detailed for instance in [55].We have defined random versions of the 2-microlocal spaces, and given almost sure conditions for continuous processes to belong to such spaces. More precise results have also been obtained for Gaussian processes. A preliminary investigation of the 2-microlocal behaviour of Wiener integrals has been performed.

Multifractal analysis of stochastic processes

A direct use of the local regularity is often fruitful in applications. This is for instance the case in RR analysis or terrain modeling. However, in some situations, it is interesting to supplement or replace it by a more global approach known as *multifractal analysis* (MA). The idea behind MA is to group together all points with same regularity (as measured by the pointwise Hölder exponent) and to measure the "size" of the sets thus obtained [28], [47], [50]. There are mainly two ways to do so, a geometrical and a statistical one.

In the geometrical approach, one defines the *Hausdorff multifractal spectrum* of a process or function X as the function: $\alpha\mapsto f_h(\alpha)=\dim\{t:\alpha_X(t)=\alpha\}$, where $\dim E$ denotes the Hausdorff dimension of the set E. This gives a fine measure-theoretic information, but is often difficult to compute theoretically, and almost impossible to estimate on numerical data.

The statistical path to MA is based on the so-called *large deviation multifractal spectrum*:

$$f_g(\alpha)=\lim_{\varepsilon\to 0}\liminf_{n\to\infty}\frac{\log\ N_n^\varepsilon(\alpha)}{\log\ n},$$

where:

$$N_n^\varepsilon(\alpha)=\#\{k:\alpha-\varepsilon\leq\alpha_n^k\leq\alpha+\varepsilon\},$$

and α_n^k is the "coarse grained exponent" corresponding to the interval $I_n^k=\left[\frac{k}{n},\frac{k+1}{n}\right]$, *i.e.*:

$$\alpha_n^k=\frac{\log|Y_n^k|}{-\log\ n}.$$

Here, Y_n^k is some quantity that measures the variation of X in the interval I_n^k, such as the increment, the oscillation or a wavelet coefficient.

The large deviation spectrum is typically easier to compute and to estimate than the Hausdorff one. In addition, it often gives more relevant information in applications.

Under very mild conditions (*e.g.* for instance, if the support of f_g is bounded, [27]) the concave envelope of f_g can be computed easily from an auxiliary function, called the *Legendre multifractal spectrum*. To do so, one basically interprets the spectrum f_g as a rate function in a large deviation principle (LDP): define, for $q \in \mathbb{R}$,

$$S_n(q) = \sum_{k=0}^{n-1} |Y_n^k|^q, \tag{62}$$

with the convention $0^q := 0$ for all $q \in \mathbb{R}$. Let:

$$\tau(q) = \liminf_{n\to\infty} \frac{\log S_n(q)}{-\log(n)}.$$

The Legendre multifractal spectrum of X is defined as the Legendre transform τ^* of τ:

$$f_l(\alpha) := \tau^*(\alpha) := \inf_{q\in\mathbb{R}} (q\alpha - \tau(q)).$$

To see the relation between f_g and f_l, define the sequence of random variables $Z_n := \log |Y_n^k|$ where the randomness is through a choice of k uniformly in $\{0, ..., n-1\}$. Consider the corresponding moment generating functions:

$$c_n(q) := -\frac{\log E_n[\exp(qZ_n)]}{\log(n)}$$

where E_n denotes expectation with respect to P_n, the uniform distribution on $\{0, ..., n-1\}$. A version of Gärtner-Ellis theorem ensures that if $\lim c_n(q)$ exists (in which case it equals $1 + \tau(q)$), and is differentiable, then $c^* = f_g - 1$. In this case, one says that the *weak multifractal formalism* holds, *i.e.* $f_g = f_l$. In favorable cases, this also coincides with f_h, a situation referred to as the *strong multifractal formalism*.

Multifractal spectra subsume a lot of information about the distribution of the regularity, that has proved useful in various situations. A most notable example is the strong correlation reported recently in several works between the narrowing of the multifractal spectrum of ECG and certain pathologies of the heart [51], [52]. Let us also mention the multifractality of TCP traffic, that has been both observed experimentally and proved on simplified models of TCP [2], [44].

Another colour in local regularity: jumps

As noted above, apart from Hölder exponents and their generalizations, at least another type of irregularity may sometimes be observed on certain real phenomena: discontinuities, which occur for instance on financial logs and certain biomedical signals. In this frame, it is of interest to supplement Hölder exponents and their extensions with (at least) an additional index that measures the local intensity and size of jumps. This is a topic we intend to pursue in full generality in the near future. So far, we have developed an approach in the particular frame of *multistable processes*. We refer to section 3.3 for more details.

3.3. Stochastic models

The second axis in the theoretical developments of the *Regularity* team aims at defining and studying stochastic processes for which various aspects of the local regularity may be prescribed.

Multifractional Brownian motion

One of the simplest stochastic process for which some kind of control over the Hölder exponents is possible is probably fractional Brownian motion (fBm). This process was defined by Kolmogorov and further studied by Mandelbrot and Van Ness, followed by many authors. The so-called "moving average" definition of fBm reads as follows:

$$Y_t = \int_{-\infty}^{0} \left[(t-u)^{H-\frac{1}{2}} - (-u)^{H-\frac{1}{2}} \right] .\mathbb{W}(du) + \int_0^t (t-u)^{H-\frac{1}{2}} .\mathbb{W}(du),$$

where $\mathbb{W}$ denotes the real white noise. The parameter H ranges in $(0,1)$, and it governs the pointwise regularity: indeed, almost surely, at each point, both the local and pointwise Hölder exponents are equal to H.

Although varying H yields processes with different regularity, the fact that the exponents are constant along any single path is often a major drawback for the modeling of real world phenomena. For instance, fBm has often been used for the synthesis natural terrains. This is not satisfactory since it yields images lacking crucial features of real mountains, where some parts are smoother than others, due, for instance, to erosion.

It is possible to generalize fBm to obtain a Gaussian process for which the pointwise Hölder exponent may be tuned at each point: the *multifractional Brownian motion (mBm)* is such an extension, obtained by substituting the constant parameter $H \in (0,1)$ with a *regularity functionH* : $\mathbb{R}_+ \to (0,1)$.

mBm was introduced independently by two groups of authors: on the one hand, Peltier and Levy-Vehel [29] defined the mBm $\{X_t;\ t \in \mathbb{R}_+\}$ from the moving average definition of the fractional Brownian motion, and set:

$$X_t = \int_{-\infty}^{0} \left[(t-u)^{H(t)-\frac{1}{2}} - (-u)^{H(t)-\frac{1}{2}} \right] .\mathbb{W}(du) + \int_0^t (t-u)^{H(t)-\frac{1}{2}} .\mathbb{W}(du),$$

On the other hand, Benassi, Jaffard and Roux [45] defined the mBm from the harmonizable representation of the fBm, *i.e.*:

$$X_t = \int_{\mathbb{R}} \frac{e^{it\xi} - 1}{|\xi|^{H(t)+\frac{1}{2}}} .\widehat{\mathbb{W}}(d\xi),$$

where $\widehat{\mathbb{W}}$ denotes the complex white noise.

The Hölder exponents of the mBm are prescribed almost surely: the pointwise Hölder exponent is $\alpha_X(t) = H(t) \wedge \alpha_H(t)$ a.s., and the local Hölder exponent is $\widetilde{\alpha}_X(t) = H(t) \wedge \widetilde{\alpha}_H(t)$ a.s. Consequently, the regularity of the sample paths of the mBm are determined by the function H or by its regularity. The multifractional Brownian motion is our prime example of a stochastic process with prescribed local regularity.

The fact that the local regularity of mBm may be tuned *via* a functional parameter has made it a useful model in various areas such as finance, biomedicine, geophysics, image analysis, ... A large number of studies have been devoted worldwide to its mathematical properties, including in particular its local time. In addition, there is now a rather strong body of work dealing the estimation of its functional parameter, *i.e.* its local regularity. See http://regularity.saclay.inria.fr/theory/stochasticmodels/bibliombm for a partial list of works, applied or theoretical, that deal with mBm.

Self-regulating processes

We have recently introduced another class of stochastic models, inspired by mBm, but where the local regularity, instead of being tuned "exogenously", is a function of the amplitude. In other words, at each point t, the Hölder exponent of the process X verifies almost surely $\alpha_X(t) = g(X(t))$, where g is a fixed deterministic function verifying certain conditions. A process satisfying such an equation is generically termed a *self-regulating process* (SRP). The particular process obtained by adapting adequately mBm is called the self-regulating multifractional process [3]. Another instance is given by modifying the Lévy construction of Brownian motion [4]. The motivation for introducing self-regulating processes is based on the following general fact: in nature, the local regularity of a phenomenon is often related to its amplitude. An intuitive example is provided by natural terrains: in young mountains, regions at higher altitudes are typically more irregular than regions at lower altitudes. We have verified this fact experimentally on several digital elevation models [8]. Other natural phenomena displaying a relation between amplitude and exponent include temperatures records and RR intervals extracted from ECG [9].

To build the SRMP, one starts from a field of fractional Brownian motions $B(t,H)$, where (t,H) span $[0,1]\times[a,b]$ and $0<a<b<1$. For each fixed H, $B(t,H)$ is a fractional Brownian motion with exponent H. Denote:

$$\overline{\underline{X}}^{\beta'}_{\alpha'} = \alpha' + (\beta'-\alpha')\frac{X-\min_K(X)}{\max_K(X)-\min_K(X)}$$

the affine rescaling between α' and β' of an arbitrary continuous random field over a compact set K. One considers the following (stochastic) operator, defined almost surely:

$$\begin{aligned}\Lambda_{\alpha',\beta'} : \mathcal{C}([0,1],[\alpha,\beta]) &\to \mathcal{C}([0,1],[\alpha,\beta])\\ Z(.) &\mapsto \overline{\underline{B(.,g(Z(.))}}^{\beta'}_{\alpha'}\end{aligned}$$

where $\alpha \le \alpha' < \beta' \le \beta$, α and β are two real numbers, and α', β' are random variables adequately chosen. One may show that this operator is contractive with respect to the sup-norm. Its unique fixed point is the SRMP. Additional arguments allow to prove that, indeed, the Hölder exponent at each point is almost surely $g(t)$.

An example of a two dimensional SRMP with function $g(x) = 1 - x^2$ is displayed on figure 1 .

We believe that SRP open a whole new and very promising area of research.

Multistable processes

Non-continuous phenomena are commonly encountered in real-world applications, *e.g.* financial records or EEG traces. For such processes, the information brought by the Hölder exponent must be supplemented by some measure of the density and size of jumps. Stochastic processes with jumps, and in particular Lévy processes, are currently an active area of research.

The simplest class of non-continuous Lévy processes is maybe the one of stable processes [56]. These are mainly characterized by a parameter $\alpha \in (0,2]$, the *stability index* ($\alpha = 2$ corresponds to the Gaussian case, that we do not consider here). This index measures in some precise sense the intensity of jumps. Paths of stable processes with α close to 2 tend to display "small jumps", while, when α is near 0, their aspect is governed by large ones.

In line with our quest for the characterization and modeling of various notions of local regularity, we have defined *multistable processes*. These are processes which are "locally" stable, but where the stability index α is now a function of time. This allows to model phenomena which, at times, are "almost continuous", and at others display large discontinuities. Such a behaviour is for instance obvious on almost any sufficiently long financial record.

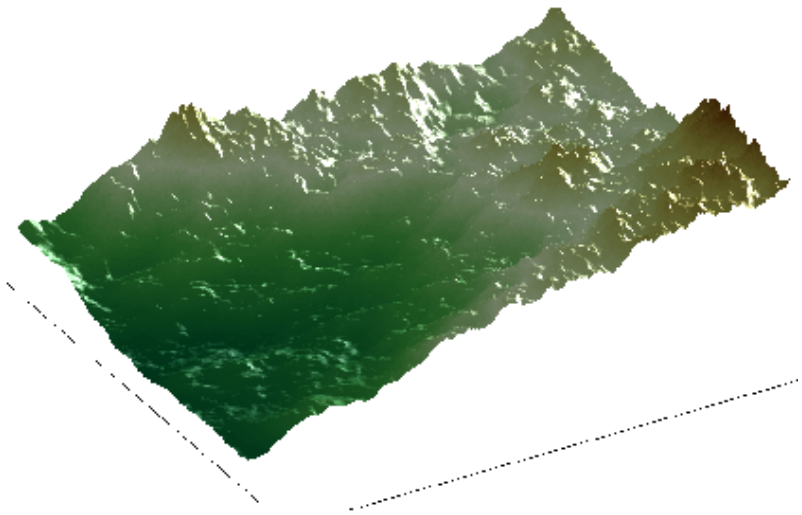


Figure 1. Self-regulating miltifractional process with $g(x) = 1 - x^2$

More formally, a multistable process is a process which is, at each time u, tangent to a stable process [49]. Recall that a process Y is said to be tangent at u to the process Y'_u if:

$$\lim_{r \to 0} \frac{Y(u+rt) - Y(u)}{r^h} = Y'_u(t), \tag{63}$$

where the limit is understood either in finite dimensional distributions or in the stronger sense of distributions. Note Y'_u may and in general will vary with u.

One approach to defining multistable processes is similar to the one developed for constructing mBm [29]: we consider fields of stochastic processes $X(t,u)$, where t is time and u is an independent parameter that controls the variation of α. We then consider a "diagonal" process $Y(t) = X(t,t)$, which will be, under certain conditions, "tangent" at each point t to a process $t \mapsto X(t,u)$.

A particular class of multistable processes, termed "linear multistable multifractional motions" (lmmm) takes the following form [11], [10]. Let $(E, \mathcal{E}, m)$ be a σ-finite measure space, and Π be a Poisson process on $E \times \mathbb{R}$ with mean measure $m \times \mathcal{L}$ ($\mathcal{L}$ denotes the Lebesgue measure). An lmmm is defined as:

$$Y(t) = a(t) \sum_{(\mathsf{X},\mathsf{Y}) \in \Pi} \mathsf{Y}^{<-1/\alpha(t)>} \left(|t - \mathsf{X}|^{h(t)-1/\alpha(t)} - |\mathsf{X}|^{h(t)-1/\alpha(t)} \right) \quad (t \in \mathbb{R}). \tag{64}$$

where $x^{<y>} := \text{sign}(x)|x|^y$, $a : \mathbb{R} \to \mathbb{R}^+$ is a C^1 function and $\alpha : \mathbb{R} \to (0,2)$ and $h : \mathbb{R} \to (0,1)$ are C^2 functions.

In fact, lmmm are somewhat more general than said above: indeed, the couple (h, α) allows to prescribe at each point, under certain conditions, both the pointwise Hölder exponent and the local intensity of jumps. In this sense, they generalize both the mBm and the linear multifractional stable motion [57]. From a broader perspective, such multistable multifractional processes are expected to provide relevant models for TCP traces, financial logs, EEG and other phenomena displaying time-varying regularity both in terms of Hölder exponents and discontinuity structure.

Figure 2 displays a graph of an lmmm with linearly increasing α and linearly decreasing H. One sees that the path has large jumps at the beginning, and almost no jumps at the end. Conversely, it is smooth (between jumps) at the beginning, but becomes jaggier and jaggier as time evolves.

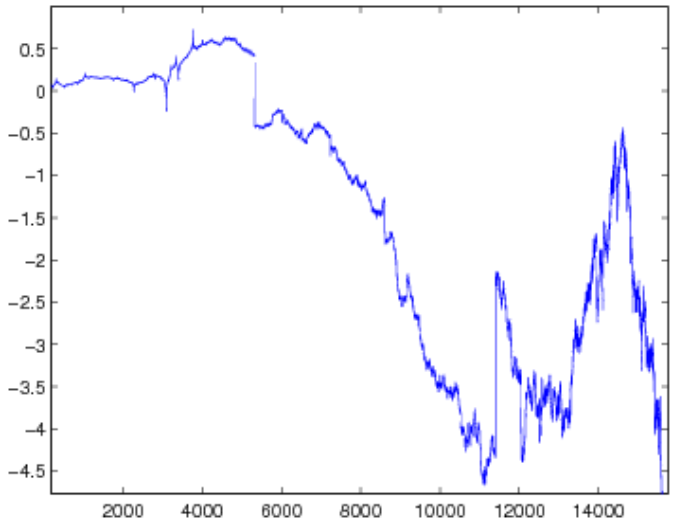


Figure 2. Linear multistable multifractional motion with linearly increasing α and linearly decreasing H

SELECT Project-Team

3. Research Program

3.1. General presentation

We learned from the applications we treated that some assumptions which are currently used in asymptotic theory for model selection are often irrelevant in practice. For instance, it is not realistic to assume that the target belongs to the family of models in competition. Moreover, in many situations, it is useful to make the size of the model depend on the sample size which make the asymptotic analysis breakdown. An important aim of SELECT is to propose model selection criteria which take these practical constraints into account.

3.2. A non asymptotic view for model selection

An important purpose of SELECT is to build and analyze penalized log-likelihood model selection criteria that are efficient when the number of models in competition grows to infinity with the number of observations. Concentration inequalities are a key tool for that purpose and lead to data-driven penalty choice strategies. A major issue of SELECT consists of deepening the analysis of data-driven penalties both from the theoretical and the practical side. There is no universal way of calibrating penalties but there are several different general ideas that we want to develop, including heuristics derived from the Gaussian theory, special strategies for variable selection and using resampling methods.

3.3. Taking into account the modeling purpose in model selection

Choosing a model is not only difficult theoretically. From a practical point of view, it is important to design model selection criteria that accommodate situations in which the data probability distribution P is unknown and which take the model user's purpose into account. Most standard model selection criteria assume that P belongs to one of a set of models, without considering the purpose of the model. By also considering the model user's purpose, we avoid or overcome certain theoretical difficulties and can produce flexible model selection criteria with data-driven penalties. The latter is useful in supervised Classification and hidden-structure models.

3.4. Bayesian model selection

The Bayesian approach to statistical problems is fundamentally probabilistic. A joint probability distribution is used to describe the relationships among all the unknowns and the data. Inference is then based on the posterior distribution i.e. the conditional probability distribution of the parameters given the observed data. Exploiting the internal consistency of the probability framework, the posterior distribution extracts the relevant information in the data and provides a complete and coherent summary of post-data uncertainty. Using the posterior to solve specific inference and decision problems is then straightforward, at least in principle.

SEQUEL Project-Team

3. Research Program

3.1. In Short

SEQUEL is primarily grounded on two domains:

- the problem of decision under uncertainty,
- statistical analysis and statistical learning, which provide the general concepts and tools to solve this problem.

To help the reader who is unfamiliar with these questions, we briefly present key ideas below.

3.2. Decision-making Under Uncertainty

The phrase "Decision under uncertainty" refers to the problem of taking decisions when we do not have a full knowledge neither of the situation, nor of the consequences of the decisions, as well as when the consequences of decision are non deterministic.

We introduce two specific sub-domains, namely the Markov decision processes which models sequential decision problems, and bandit problems.

3.2.1. Reinforcement Learning

Sequential decision processes occupy the heart of the SEQUEL project; a detailed presentation of this problem may be found in Puterman's book [48].

A Markov Decision Process (MDP) is defined as the tuple $(\mathcal{X}, \mathcal{A}, P, r)$ where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, P is the probabilistic transition kernel, and $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is the reward function. For the sake of simplicity, we assume in this introduction that the state and action spaces are finite. If the current state (at time t) is $x \in \mathcal{X}$ and the chosen action is $a \in \mathcal{A}$, then the Markov assumption means that the transition probability to a new state $x' \in \mathcal{X}$ (at time $t+1$) only depends on (x, a). We write $p(x'|x, a)$ the corresponding transition probability. During a transition $(x, a) \to x'$, a reward $r(x, a, x')$ is incurred.

In the MDP $(\mathcal{X}, \mathcal{A}, P, r)$, each initial state x_0 and action sequence $a_0, a_1, ...$ gives rise to a sequence of states $x_1, x_2, ...$, satisfying $\mathbb{P}(x_{t+1} = x'|x_t = x, a_t = a) = p(x'|x, a)$, and rewards [0]$r_1, r_2, ...$ defined by $r_t = r(x_t, a_t, x_{t+1})$.

The history of the process up to time t is defined to be $H_t = (x_0, a_0, ..., x_{t-1}, a_{t-1}, x_t)$. A policy π is a sequence of functions $\pi_0, \pi_1, ...$, where π_t maps the space of possible histories at time t to the space of probability distributions over the space of actions $\mathcal{A}$. To follow a policy means that, in each time step, we assume that the process history up to time t is $x_0, a_0, ..., x_t$ and the probability of selecting an action a is equal to $\pi_t(x_0, a_0, ..., x_t)(a)$. A policy is called stationary (or Markovian) if π_t depends only on the last visited state. In other words, a policy $\pi = (\pi_0, \pi_1, ...)$ is called stationary if $\pi_t(x_0, a_0, ..., x_t) = \pi_0(x_t)$ holds for all $t \geq 0$. A policy is called deterministic if the probability distribution prescribed by the policy for any history is concentrated on a single action. Otherwise it is called a stochastic policy.

[0]Note that for simplicity, we considered the case of a deterministic reward function, but in many applications, the reward r_t itself is a random variable.

We move from an MD process to an MD problem by formulating the goal of the agent, that is what the sought policy π has to optimize? It is very often formulated as maximizing (or minimizing), in expectation, some functional of the sequence of future rewards. For example, an usual functional is the infinite-time horizon sum of discounted rewards. For a given (stationary) policy π, we define the value function $V^\pi(x)$ of that policy π at a state $x \in \mathcal{X}$ as the expected sum of discounted future rewards given that we state from the initial state x and follow the policy π:

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = x, \pi\right], \tag{65}$$

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0,1)$ is the discount factor. This value function V^π gives an evaluation of the performance of a given policy π. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [43]) and average reward settings. Note also that, here, we considered the problem of maximizing a reward functional, but a formulation in terms of minimizing some cost or risk functional would be equivalent.

In order to maximize a given functional in a sequential framework, one usually applies Dynamic Programming (DP) [41], which introduces the optimal value function $V^*(x)$, defined as the optimal expected sum of rewards when the agent starts from a state x. We have $V^*(x) = \sup_\pi V^\pi(x)$. Now, let us give two definitions about policies:

- We say that a policy π is optimal, if it attains the optimal values $V^*(x)$ for any state $x \in \mathcal{X}$, *i.e.*, if $V^\pi(x) = V^*(x)$ for all $x \in \mathcal{X}$. Under mild conditions, deterministic stationary optimal policies exist [42]. Such an optimal policy is written π^*.
- We say that a (deterministic stationary) policy π is greedy with respect to (w.r.t.) some function V (defined on $\mathcal{X}$) if, for all $x \in \mathcal{X}$,

$$\pi(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V(x')\right].$$

  where $\arg\max_{a \in \mathcal{A}} f(a)$ is the set of $a \in \mathcal{A}$ that maximizes $f(a)$. For any function V, such a greedy policy always exists because $\mathcal{A}$ is finite.

The goal of Reinforcement Learning (RL), as well as that of dynamic programming, is to design an optimal policy (or a good approximation of it).

The well-known Dynamic Programming equation (also called the Bellman equation) provides a relation between the optimal value function at a state x and the optimal value function at the successors states x' when choosing an optimal action: for all $x \in \mathcal{X}$,

$$V^*(x) = \max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{66}$$

The benefit of introducing this concept of optimal value function relies on the property that, from the optimal value function V^*, it is easy to derive an optimal behavior by choosing the actions according to a policy greedy w.r.t. V^*. Indeed, we have the property that a policy greedy w.r.t. the optimal value function is an optimal policy:

$$\pi^*(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{67}$$

In short, we would like to mention that most of the reinforcement learning methods developed so far are built on one (or both) of the two following approaches ([54]):

- Bellman's dynamic programming approach, based on the introduction of the value function. It consists in learning a "good" approximation of the optimal value function, and then using it to derive a greedy policy w.r.t. this approximation. The hope (well justified in several cases) is that the performance V^π of the policy π greedy w.r.t. an approximation V of V^* will be close to optimality. This approximation issue of the optimal value function is one of the major challenges inherent to the reinforcement learning problem. **Approximate dynamic programming** addresses the problem of estimating performance bounds (*e.g.* the loss in performance $||V^* - V^\pi||$ resulting from using a policy π-greedy w.r.t. some approximation V- instead of an optimal policy) in terms of the approximation error $||V^* - V||$ of the optimal value function V^* by V. Approximation theory and Statistical Learning theory provide us with bounds in terms of the number of sample data used to represent the functions, and the capacity and approximation power of the considered function spaces.
- Pontryagin's maximum principle approach, based on sensitivity analysis of the performance measure w.r.t. some control parameters. This approach, also called **direct policy search** in the Reinforcement Learning community aims at directly finding a good feedback control law in a parameterized policy space without trying to approximate the value function. The method consists in estimating the so-called **policy gradient**, *i.e.* the sensitivity of the performance measure (the value function) w.r.t. some parameters of the current policy. The idea being that an optimal control problem is replaced by a parametric optimization problem in the space of parameterized policies. As such, deriving a policy gradient estimate would lead to performing a stochastic gradient method in order to search for a local optimal parametric policy.

Finally, many extensions of the Markov decision processes exist, among which the Partially Observable MDPs (POMDPs) is the case where the current state does not contain all the necessary information required to decide for sure of the best action.

3.2.2. Multi-arm Bandit Theory

Bandit problems illustrate the fundamental difficulty of decision making in the face of uncertainty: A decision maker must choose between what seems to be the best choice ("exploit"), or to test ("explore") some alternative, hoping to discover a choice that beats the current best choice.

The classical example of a bandit problem is deciding what treatment to give each patient in a clinical trial when the effectiveness of the treatments are initially unknown and the patients arrive sequentially. These bandit problems became popular with the seminal paper [49], after which they have found applications in diverse fields, such as control, economics, statistics, or learning theory.

Formally, a K-armed bandit problem ($K \geq 2$) is specified by K real-valued distributions. In each time step a decision maker can select one of the distributions to obtain a sample from it. The samples obtained are considered as rewards. The distributions are initially unknown to the decision maker, whose goal is to maximize the sum of the rewards received, or equivalently, to minimize the regret which is defined as the loss compared to the total payoff that can be achieved given full knowledge of the problem, *i.e.*, when the arm giving the highest expected reward is pulled all the time.

The name "bandit" comes from imagining a gambler playing with K slot machines. The gambler can pull the arm of any of the machines, which produces a random payoff as a result: When arm k is pulled, the random payoff is drawn from the distribution associated to k. Since the payoff distributions are initially unknown, the gambler must use exploratory actions to learn the utility of the individual arms. However, exploration has to be carefully controlled since excessive exploration may lead to unnecessary losses. Hence, to play well, the gambler must carefully balance exploration and exploitation. Auer *et al.* [40] introduced the algorithm UCB (Upper Confidence Bounds) that follows what is now called the "optimism in the face of uncertainty principle". Their algorithm works by computing upper confidence bounds for all the arms and then choosing the arm with the highest such bound. They proved that the expected regret of their algorithm increases at most

at a logarithmic rate with the number of trials, and that the algorithm achieves the smallest possible regret up to some sub-logarithmic factor (for the considered family of distributions).

3.3. Statistical analysis of time series

Many of the problems of machine learning can be seen as extensions of classical problems of mathematical statistics to their (extremely) non-parametric and model-free cases. Other machine learning problems are founded on such statistical problems. Statistical problems of sequential learning are mainly those that are concerned with the analysis of time series. These problems are as follows.

3.3.1. Prediction of Sequences of Structured and Unstructured Data

Given a series of observations $x_1, \cdots, x_n$ it is required to give forecasts concerning the distribution of the future observations $x_{n+1}, x_{n+2}, \cdots$; in the simplest case, that of the next outcome x_{n+1}. Then x_{n+1} is revealed and the process continues. Different goals can be formulated in this setting. One can either make some assumptions on the probability measure that generates the sequence $x_1, \cdots, x_n, \cdots$, such as that the outcomes are independent and identically distributed (i.i.d.), or that the sequence is a Markov chain, that it is a stationary process, etc. More generally, one can assume that the data is generated by a probability measure that belongs to a certain set $\mathcal{C}$. In these cases the goal is to have the discrepancy between the predicted and the "true" probabilities to go to zero, if possible, with guarantees on the speed of convergence.

Alternatively, rather than making some assumptions on the data, one can change the goal: the predicted probabilities should be asymptotically as good as those given by the best reference predictor from a certain pre-defined set.

Another dimension of complexity in this problem concerns the nature of observations x_i. In the simplest case, they come from a finite space, but already basic applications often require real-valued observations. Moreover, function or even graph-valued observations often arise in practice, in particular in applications concerning Web data. In these settings estimating even simple characteristics of probability distributions of the future outcomes becomes non-trivial, and new learning algorithms for solving these problems are in order.

3.3.2. Hypothesis testing

Given a series of observations of $x_1, \cdots, x_n, \cdots$ generated by some unknown probability measure μ, the problem is to test a certain given hypothesis H_0 about μ, versus a given alternative hypothesis H_1. There are many different examples of this problem. Perhaps the simplest one is testing a simple hypothesis "μ is Bernoulli i.i.d. measure with probability of 0 equals 1/2" versus "μ is Bernoulli i.i.d. with the parameter different from 1/2". More interesting cases include the problems of model verification: for example, testing that μ is a Markov chain, versus that it is a stationary ergodic process but not a Markov chain. In the case when we have not one but several series of observations, we may wish to test the hypothesis that they are independent, or that they are generated by the same distribution. Applications of these problems to a more general class of machine learning tasks include the problem of feature selection, the problem of testing that a certain behaviour (such as pulling a certain arm of a bandit, or using a certain policy) is better (in terms of achieving some goal, or collecting some rewards) than another behaviour, or than a class of other behaviours.

The problem of hypothesis testing can also be studied in its general formulations: given two (abstract) hypothesis H_0 and H_1 about the unknown measure that generates the data, find out whether it is possible to test H_0 against H_1 (with confidence), and if yes then how can one do it.

3.3.3. Change Point Analysis

A stochastic process is generating the data. At some point, the process distribution changes. In the "offline" situation, the statistician observes the resulting sequence of outcomes and has to estimate the point or the points at which the change(s) occurred. In online setting, the goal is to detect the change as quickly as possible.

These are the classical problems in mathematical statistics, and probably among the last remaining statistical problems not adequately addressed by machine learning methods. The reason for the latter is perhaps in that the problem is rather challenging. Thus, most methods available so far are parametric methods concerning piece-wise constant distributions, and the change in distribution is associated with the change in the mean. However, many applications, including DNA analysis, the analysis of (user) behaviour data, etc., fail to comply with this kind of assumptions. Thus, our goal here is to provide completely non-parametric methods allowing for any kind of changes in the time-series distribution.

3.3.4. Clustering Time Series, Online and Offline

The problem of clustering, while being a classical problem of mathematical statistics, belongs to the realm of unsupervised learning. For time series, this problem can be formulated as follows: given several samples $x^1 = (x_1^1, \cdots, x_{n_1}^1), \cdots, x^N = (x_N^1, \cdots, x_{n_N}^N)$, we wish to group similar objects together. While this is of course not a precise formulation, it can be made precise if we assume that the samples were generated by k different distributions.

The online version of the problem allows for the number of observed time series to grow with time, in general, in an arbitrary manner.

3.3.5. Online Semi-Supervised Learning

Semi-supervised learning (SSL) is a field of machine learning that studies learning from both labeled and unlabeled examples. This learning paradigm is extremely useful for solving real-world problems, where data is often abundant but the resources to label them are limited.

Furthermore, *online* SSL is suitable for adaptive machine learning systems. In the classification case, learning is viewed as a repeated game against a potentially adversarial nature. At each step t of this game, we observe an example $\mathbf{x_t}$, and then predict its label $\widehat{y}_t$.

The challenge of the game is that we only exceptionally observe the true label y_t. In the extreme case, which we also study, only a handful of labeled examples are provided in advance and set the initial bias of the system while unlabeled examples are gathered online and update the bias continuously. Thus, if we want to adapt to changes in the environment, we have to rely on indirect forms of feedback, such as the structure of data.

3.4. Statistical Learning and Bayesian Analysis

Before detailing some issues in these fields, let us remind the definition of a few terms.

Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness.

Statistical learning is an approach to machine intelligence that is based on statistical modeling of data. With a statistical model in hand, one applies probability theory and decision theory to get an algorithm. This is opposed to using training data merely to select among different algorithms or using heuristics/"common sense" to design an algorithm.

Bayesian Analysis applies to data that could be seen as observations in the more general meaning of the term. These data may not only come from classical sensors but also from any *device* recording information. From an operational point of view, like for statistical learning, uncertainty about the data is modeled by a probability measure thus defining the so-called likelihood functions. This last one depends upon parameters defining the state of the world we focus on for decision purposes. Within the Bayesian framework the uncertainty about these parameters is also modeled by probability measures, the priors that are subjective probabilities. Using probability theory and decision theory, one then defines new algorithms to estimate the parameters of interest and/or associated decisions. According to the International Society for Bayesian Analysis (source: http://bayesian.org), and from a more general point of view, this overall process could be

summarize as follows: one assesses the current state of knowledge regarding the issue of interest, gather new data to address remaining questions, and then update and refine their understanding to incorporate both new and old data. Bayesian inference provides a logical, quantitative framework for this process based on probability theory.

Kernel method. Generally speaking, a kernel function is a function that maps a couple of points to a real value. Typically, this value is a measure of dissimilarity between the two points. Assuming a few properties on it, the kernel function implicitly defines a dot product in some function space. This very nice formal property as well as a bunch of others have ensured a strong appeal for these methods in the last 10 years in the field of function approximation. Many classical algorithms have been "kernelized", that is, restated in a much more general way than their original formulation. Kernels also implicitly induce the representation of data in a certain "suitable" space where the problem to solve (classification, regression, ...) is expected to be simpler (non-linearity turns to linearity).

The fundamental tools used in SEQUEL come from the field of statistical learning [45]. We briefly present the most important for us to date, namely, kernel-based non parametric function approximation, and non parametric Bayesian models.

3.4.1. Non-parametric methods for Function Approximation

In statistics in general, and applied mathematics, the approximation of a multi-dimensional real function given some samples is a well-known problem (known as either regression, or interpolation, or function approximation, ...). Regressing a function from data is a key ingredient of our research, or to the least, a basic component of most of our algorithms. In the context of sequential learning, we have to regress a function while data samples are being obtained one at a time, while keeping the constraint to be able to predict points at any step along the acquisition process. In sequential decision problems, we typically have to learn a value function, or a policy.

Many methods have been proposed for this purpose. We are looking for suitable ones to cope with the problems we wish to solve. In reinforcement learning, the value function may have areas where the gradient is large; these are areas where the approximation is difficult, while these are also the areas where the accuracy of the approximation should be maximal to obtain a good policy (and where, otherwise, a bad choice of action may imply catastrophic consequences).

We particularly favor non parametric methods since they make quite a few assumptions about the function to learn. In particular, we have strong interests in l_1-regularization, and the (kernelized-)LARS algorithm. l_1-regularization yields sparse solutions, and the LARS approach produces the whole regularization path very efficiently, which helps solving the regularization parameter tuning problem.

3.4.2. Nonparametric Bayesian Estimation

Numerous problems may be solved efficiently by a Bayesian approach. The use of Monte-Carlo methods allows us to handle non–linear, as well as non–Gaussian, problems. In their standard form, they require the formulation of probability densities in a parametric form. For instance, it is a common usage to use Gaussian likelihood, because it is handy. However, in some applications such as Bayesian filtering, or blind deconvolution, the choice of a parametric form of the density of the noise is often arbitrary. If this choice is wrong, it may also have dramatic consequences on the estimation quality. To overcome this shortcoming, one possible approach is to consider that this density must also be estimated from data. A general Bayesian approach then consists in defining a probabilistic space associated with the possible outcomes of the *object* to be estimated. Applied to density estimation, it means that we need to define a probability measure on the probability density of the noise: such a measure is called a *random measure*. The classical Bayesian inference procedures can then been used. This approach being by nature non parametric, the associated frame is called *Non Parametric Bayesian*.

In particular, mixtures of Dirichlet processes [44] provide a very powerful formalism. Dirichlet Processes are a possible random measure and Mixtures of Dirichlet Processes are an extension of well-known finite mixture models. Given a mixture density $f(x|\theta)$, and $G(d\theta) = \sum_{k=1}^{\infty} \omega_k \delta_{U_k}(d\theta)$, a Dirichlet process, we define a mixture of Dirichlet processes as:

$$F(x) = \int_{\Theta} f(x|\theta) G(d\theta) = \sum_{k=1}^{\infty} \omega_k f(x|U_k) \tag{68}$$

where $F(x)$ is the density to be estimated. The class of densities that may be written as a mixture of Dirichlet processes is very wide, so that they really fit a very large number of applications.

Given a set of observations, the estimation of the parameters of a mixture of Dirichlet processes is performed by way of a Monte Carlo Markov Chain (MCMC) algorithm. Dirichlet Process Mixture are also widely used in clustering problems. Once the parameters of a mixture are estimated, they can be interpreted as the parameters of a specific cluster defining a class as well. Dirichlet processes are well known within the machine learning community and their potential in statistical signal processing still need to be developed.

3.4.3. Random Finite Sets for multisensor multitarget tracking

In the general multi-sensor multi-target Bayesian framework, an unknown (and possibly varying) number of targets whose states $x_1, ...x_n$ are observed by several sensors which produce a collection of measurements $z_1, ..., z_m$ at every time step k. Well-known models to this problem are track-based models, such as the joint probability data association (JPDA), or joint multi-target probabilities, such as the joint multi-target probability density. Common difficulties in multi-target tracking arise from the fact that the system state and the collection of measures from sensors are unordered and their size evolve randomly through time. Vector-based algorithms must therefore account for state coordinates exchanges and missing data within an unknown time interval. Although this approach is very popular and has resulted in many algorithms in the past, it may not be the optimal way to tackle the problem, since the sate and the data are in fact *sets* and not vectors.

The random finite set theory provides a powerful framework to deal with these issues. Mahler's work on finite sets statistics (FISST) provides a mathematical framework to build multi-object densities and derive the Bayesian rules for state prediction and state estimation. Randomness on object number and their states are encapsulated into random finite sets (RFS), namely multi-target(state) sets $X = \{x_1, ..., x_n\}$ and multi-sensor (measurement) set $Zk = \{z_1, ..., z_m\}$. The objective is then to propagate the multitarget probability density $f_{k|k}(X|Z(k))$ by using the Bayesian set equations at every time step k:

$$\begin{aligned} f_{k+1|k}(X|Z^{(k)}) &= \int f_{k+1|k}(X|W) f_{k|k}(W|Z^{(k)}) \delta W \\ f_{k+1|k+1}(X|Z^{(k+1)}) &= \frac{f_{k+1}(Z_{k+1}|X) f_{k+1|k}(X|Z^{(k)})}{\int f_{k+1}(Z_{k+1}|W) f_{k+1|k}(W|Z^{(k)}) \delta W} \end{aligned} \tag{69}$$

where:

- $X = \{x_1, ..., x_n\}$ is a multi-target state, *i.e.* a finite set of elements x_i defined on the single-target space $\mathcal{X}$; [0]
- $Z_{k+1} = \{z_1, ..., z_m\}$ is the current multi-sensor observation, *i.e.* a collection of measures z_i produced at time $k+1$ by all the sensors;
- $Z^{(k)} = \bigcup_{t \leqslant k} Z_t$ is the collection of observations up to time k;
- $f_{k|k}(W|Z^{(k)})$ is the current multi-target posterior density in state W;
- $f_{k+1|k}(X|W)$ is the current multi-target Markov transition density, from state W to state X;
- $f_{k+1}(Z|X)$ is the current multi-sensor/multi-target likelihood function.

[0] The state x_i of a target is usually composed of its position, its velocity, etc.

Although equations (5) may seem similar to the classical single-sensor/single-target Bayesian equations, they are generally intractable because of the presence of the *set integrals*. For, a RFS Ξ is characterized by the family of its Janossy densities $j_{\Xi,1}(x_1)$, $j_{\Xi,2}(x_1,x_2)$... and not just by one density as it is the case with vectors. Mahler then introduced the PHD, defined on single-target state space. The PHD is the quantity whose integral on any region S is the expected number of targets inside S. Mahler proved that the PHD is the first-moment density of the multi-target probability density. Although defined on single-state space X, the PHD encapsulates information on both target number and states.

SIERRA Project-Team

3. Research Program

3.1. Supervised Learning

This part of our research focuses on methods where, given a set of examples of input/output pairs, the goal is to predict the output for a new input, with research on kernel methods, calibration methods, and multi-task learning.

3.2. Unsupervised Learning

We focus here on methods where no output is given and the goal is to find structure of certain known types (e.g., discrete or low-dimensional) in the data, with a focus on matrix factorization, statistical tests, dimension reduction, and semi-supervised learning.

3.3. Parsimony

The concept of parsimony is central to many areas of science. In the context of statistical machine learning, this takes the form of variable or feature selection. The team focuses primarily on structured sparsity, with theoretical and algorithmic contributions (this is the main topic of the ERC starting investigator grant awarded to F. Bach).

3.4. Optimization

Optimization in all its forms is central to machine learning, as many of its theoretical frameworks are based at least in part on empirical risk minimization. The team focuses primarily on convex and bandit optimization, with a particular focus on large-scale optimization.

TAO Project-Team

3. Research Program

3.1. The Four Pillars of TAO

This Section describes TAO main research directions at the crossroad of Machine Learning and Evolutionary Computation. Since 2008, TAO has been structured in several special interest groups (SIGs) to enable the agile investigation of long-term or emerging theoretical and applicative issues. The comparatively small size of TAO SIGs enables in-depth and lively discussions; the fact that all TAO members belong to several SIGs, on the basis of their personal interests, enforces the strong and informal collaboration of the groups, and the fast information dissemination.

The first two SIGs consolidate the key TAO scientific pillars, while the others evolve and adapt to new topics.

The **Stochastic Continuous Optimization** SIG (OPT-SIG) takes advantage of the fact that TAO is acknowledged the best French research group and one of the top international groups in evolutionary computation from a theoretical and algorithmic standpoint. A main priority on the OPT-SIG research agenda is to provide theoretical and algorithmic guarantees for the current world state-of-the-art continuous stochastic optimizer, CMA-ES, ranging from convergence analysis (Youhei Akimoto's post-docs) to a rigorous benchmarking methodology. Incidentally, this benchmark platform COCO has been acknowledged since 2009 as "the" international continuous optimization benchmark, and its extension is at the core of the ANR project NumBBO (started end 2012). Another priority is to address the current limitations of CMA-ES in terms of high-dimensional or expensive optimization and constraint handling (respectively Ouassim Ait El Hara's, Ilya Loshchilov's PhDs and Asma Atamna's).

The **Optimal Decision Making under Uncertainty** SIG (UCT-SIG) benefits from the MoGo expertise (see Section 5.2 and the team previous activity reports) and its past and present world records in the domain of computer-Go, establishing the international visibility of TAO in sequential decision making. Since 2010, UCT-SIG resolutely moves to address the problems of energy management from a fundamental and applied perspective. On the one hand, energy management offers a host of challenging issues, ranging from long-horizon policy optimization to the combinatorial nature of the search space, from the modeling of prior knowledge to non-stationary environment to name a few. On the other hand, the energy management issue can hardly be tackled in a pure academic perspective: tight collaborations with industrial partners are needed to access the true operational constraints. Such international and national collaborations have been started by Olivier Teytaud during his three stays (1 year, 6 months, 6 months) in Taiwan, and witnessed by the FP7 STREP Citines, the ADEME Post contract, and the METIS I-lab with SME Artelys.

The **E-Science** SIG (E-S-SIG) replaces and extends the former *Distributed systems* SIG, that was devoted to the modeling and optimization of (large scale) distributed systems, and itself was extending the goals of the original *Autonomic Computing* SIG, initiated by Cécile Germain-Renaud and investigating the use of statistical Machine Learning for large scale computational architectures, from data acquisition (the Grid Observatory in the European Grid Initiative) to grid management and fault detection. Indeed, how to model and manage network-based activities has been acknowledged a key topic *per se*, including the modeling of multi-agent systems and the exploitation of simulation results in the SimTools RNSC network frame. Further extensions are still being developped in the context of the TIMCO FUI project (started end 2012); the challenge is not only to port ML algorithms on massively distributed architectures, but to see how these architectures can inspire new ML criteria and methodologies. But these activities have become more and more application-driven, from High Energy Physics for the highly distributed computation to the Social Sciences for the multi-agents approaches – hence the change of focus of this SIG. A major result of this theme is the creation of the Paris-Saclay Center for Data Science, co-chaired by Balázs Kégl, and the organization of the Higgs-ML challenge (http://higgsml.lal.in2p3.fr/), most popular challenge ever on the Kaggle platform.

The **Designing Criteria** SIG (CRI-SIG) focuses on the design of learning and optimization criteria. It elaborates on the lessons learned from the former *Complex Systems* SIG, showing that the key issue in challenging applications often is to design the objective itself. Such targeted criteria are pervasive in the study and building of autonomous cognitive systems, ranging from intrinsic rewards in robotics to the notion of saliency in vision and image understanding. The desired criteria can also result from fundamental requirements, such as scale invariance in a statistical physics perspective, and guide the algorithmic design. Additionally, the criteria can also be domain-driven and reflect the expert priors concerning the structure of the sought solution (e.g., spatio-temporal consistency); the challenge is to formulate such criteria in a mixed convex/non differentiable objective function, amenable to tractable optimization.

The activity of the former *Crossing the Chasm* SIG gradually decreased after the completion of the 2 PhD theses funded by the Microsoft/Inria joint lab (Adapt project) and devoted to hyper-parameter tuning. As a matter of fact, though not a major research topic any more, hyper-parameter tuning has become pervasive in TAO, chiefly for continuous optimization (OPT-SIG, Section 6.3), AI planning (CRI-SIG, Section 6.5) and Air Traffic Control Optimization (Section 4.2). Recent work addressing algorithm selection using Collaborative Filtering algorithms (CRI-SIG, Section 6.5) can (and will) indeed be applied to hyper-parameter tuning for optimization algorithms.

TOSCA Project-Team

3. Research Program

3.1. Research Program

Most often physicists, economists, biologists, engineers need a stochastic model because they cannot describe the physical, economical, biological, etc., experiment under consideration with deterministic systems, either because of its complexity and/or its dimension or because precise measurements are impossible. Then they abandon trying to get the exact description of the state of the system at future times given its initial conditions, and try instead to get a statistical description of the evolution of the system. For example, they desire to compute occurrence probabilities for critical events such as the overstepping of a given thresholds by financial losses or neuronal electrical potentials, or to compute the mean value of the time of occurrence of interesting events such as the fragmentation to a very small size of a large proportion of a given population of particles. By nature such problems lead to complex modelling issues: one has to choose appropriate stochastic models, which require a thorough knowledge of their qualitative properties, and then one has to calibrate them, which requires specific statistical methods to face the lack of data or the inaccuracy of these data. In addition, having chosen a family of models and computed the desired statistics, one has to evaluate the sensitivity of the results to the unavoidable model specifications. The TOSCA team, in collaboration with specialists of the relevant fields, develops theoretical studies of stochastic models, calibration procedures, and sensitivity analysis methods.

In view of the complexity of the experiments, and thus of the stochastic models, one cannot expect to use closed form solutions of simple equations in order to compute the desired statistics. Often one even has no other representation than the probabilistic definition (e.g., this is the case when one is interested in the quantiles of the probability law of the possible losses of financial portfolios). Consequently the practitioners need Monte Carlo methods combined with simulations of stochastic models. As the models cannot be simulated exactly, they also need approximation methods which can be efficiently used on computers. The TOSCA team develops mathematical studies and numerical experiments in order to determine the global accuracy and the global efficiency of such algorithms.

The simulation of stochastic processes is not motivated by stochastic models only. The stochastic differential calculus allows one to represent solutions of certain deterministic partial differential equations in terms of probability distributions of functionals of appropriate stochastic processes. For example, elliptic and parabolic linear equations are related to classical stochastic differential equations, whereas nonlinear equations such as the Burgers and the Navier–Stokes equations are related to McKean stochastic differential equations describing the asymptotic behavior of stochastic particle systems. In view of such probabilistic representations one can get numerical approximations by using discretization methods of the stochastic differential systems under consideration. These methods may be more efficient than deterministic methods when the space dimension of the PDE is large or when the viscosity is small. The TOSCA team develops new probabilistic representations in order to propose probabilistic numerical methods for equations such as conservation law equations, kinetic equations, and nonlinear Fokker–Planck equations.

ABS Project-Team

3. Research Program

3.1. Introduction

The research conducted by ABS focuses on three main directions in Computational Structural Biology (CSB), together with the associated methodological developments:
– Modeling interfaces and contacts,
– Modeling macro-molecular assemblies,
– Modeling the flexibility of macro-molecules,
– Algorithmic foundations.

3.2. Modeling Interfaces and Contacts

Keywords: Docking, interfaces, protein complexes, structural alphabets, scoring functions, Voronoi diagrams, arrangements of balls.

The Protein Data Bank, http://www.rcsb.org/pdb, contains the structural data which have been resolved experimentally. Most of the entries of the PDB feature isolated proteins [0], the remaining ones being protein - protein or protein - drug complexes. These structures feature what Nature does – up to the bias imposed by the experimental conditions inherent to structure elucidation, and are of special interest to investigate non-covalent contacts in biological complexes. More precisely, given two proteins defining a complex, interface atoms are defined as the atoms of one protein *interacting* with atoms of the second one. Understanding the structure of interfaces is central to understand biological complexes and thus the function of biological molecules [39]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [42]. Current investigations follow two routes. From the experimental perspective [25], directed mutagenesis enables one to quantify the energetic importance of residues, important residues being termed *hot* residues. Such studies recently evidenced the *modular* architecture of interfaces [36]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [31].

The description of interfaces is also of special interest to improve *scoring functions*. By scoring function, two things are meant: either a function which assigns to a complex a quantity homogeneous to a free energy change [0], or a function stating that a complex is more stable than another one, in which case the value returned is a score and not an energy. Borrowing to statistical mechanics [20], the usual way to design scoring functions is to mimic the so-called potentials of mean force. To put it briefly, one reverts Boltzmann's law, that is, denoting $p_i(r)$ the probability of two atoms –defining type i– to be located at distance r, the (free) energy assigned to the pair is computed as $E_i(r) = -kT \log p_i(r)$. Estimating from the PDB one function $p_i(r)$ for each type of pair of atoms, the energy of a complex is computed as the sum of the energies of the pairs located within a distance threshold [40], [27]. To compare the energy thus obtained to a reference state, one may compute $E = \sum_i p_i \log p_i/q_i$, with p_i the observed frequencies, and q_i the frequencies stemming from an a priori model [32]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

Describing interfaces poses problems in two settings: static and dynamic.

[0] For structures resolved by crystallography, the PDB contains the asymmetric unit of the crystal. Determining the biological unit from the asymmetric unit is a problem in itself.

[0] The Gibbs free energy of a system is defined by $G = H - TS$, with $H = U + PV$. G is minimum at an equilibrium, and differences in G drive chemical reactions.

In the static setting, one seeks the minimalist geometric model providing a relevant bio-physical signal. A first step in doing so consists of identifying interface atoms, so as to relate the geometry and the bio-chemistry at the interface level [8]. To elaborate at the atomic level, one seeks a structural alphabet encoding the spatial structure of proteins. At the side-chain and backbone level, an example of such alphabet is that of [21]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [41], no such alphabet is known. Specific important local conformations are known, though. One of them is the so-called dehydron structure, which is an under-desolvated hydrogen bond – a property that can be directly inferred from the spatial configuration of the C_α carbons surrounding a hydrogen bond [24].

In the dynamic setting, one wishes to understand whether selected (hot) residues exhibit specific dynamic properties, so as to serve as anchors in a binding process [35]. More generally, any significant observation raised in the static setting deserves investigations in the dynamic setting, so as to assess its stability. Such questions are also related to the problem of correlated motions, which we discuss next.

3.3. Modeling Macro-molecular Assemblies

Keywords: Macro-molecular assembly, reconstruction by data integration, proteomics, modeling with uncertainties, curved Voronoi diagrams, topological persistence.

3.3.1. Reconstruction by Data Integration

Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [19]. In a nutshell, the strategy is reminiscent from NMR and consists of mixing experimental data from a variety of sources, so as to find out the model(s) best complying with the data. This strategy has been in particular used to propose plausible models of the Nuclear Pore Complex [18], the largest assembly known to date in the eukaryotic cell, and consisting of 456 protein *instances* of 30 *types*.

3.3.2. Modeling with Uncertainties and Model Assessment

Reconstruction by data integration requires three ingredients. First, a parametrized model must be adopted, typically a collection of balls to model a protein with pseudo-atoms. Second, as in NMR, a functional measuring the agreement between a model and the data must be chosen. In [17], this functional is based upon *restraints*, namely penalties associated to the experimental data. Third, an optimization scheme must be selected. The design of restraints is notoriously challenging, due to the ambiguous nature and/or the noise level of the data. For example, Tandem Affinity Purification (TAP) gives access to a *pullout* i.e. a list of protein types which are known to interact with one tagged protein type, but no information on the number of complexes or on the stoichiometry of proteins types within a complex is provided. In cryo-EM, the envelope enclosing an assembly is often imprecisely defined, in particular in regions of low density. For immuno-EM labelling experiments, positional uncertainties arise from the microscope resolution.

These uncertainties coupled with the complexity of the functional being optimized, which in general is non convex, have two consequences. First, it is impossible to single out a unique reconstruction, and a set of plausible reconstructions must be considered. As an example, 1000 plausible models of the NPC were reported in [17]. Interestingly, averaging the positions of all balls of a particular protein type across these models resulted in 30 so-called *probability density maps*, each such map encoding the probability of presence of a particular protein type at a particular location in the NPC. Second, the assessment of all models (individual and averaged) is non trivial. In particular, the lack of straightforward statistical analysis of the individual models and the absence of assessment for the averaged models are detrimental to the mechanistic exploitation of the reconstruction results. At this stage, such models therefore remain qualitative.

3.4. Modeling the Flexibility of Macro-molecules

Keywords: Folding, docking, energy landscapes, induced fit, molecular dynamics, conformers, conformer ensembles, point clouds, reconstruction, shape learning, Morse theory.

Proteins in vivo vibrate at various frequencies: high frequencies correspond to small amplitude deformations of chemical bonds, while low frequencies characterize more global deformations. This flexibility contributes to the entropy thus the `free energy` of the system *protein - solvent*. From the experimental standpoint, NMR studies generate ensembles of conformations, called `conformers`, and so do molecular dynamics (MD) simulations. Of particular interest while investigating flexibility is the notion of correlated motion. Intuitively, when a protein is folded, all atomic movements must be correlated, a constraint which gets alleviated when the protein unfolds since the steric constraints get relaxed [0]. Understanding correlations is of special interest to predict the folding pathway that leads a protein towards its native state. A similar discussion holds for the case of partners within a complex, for example in the third step of the *diffusion - conformer selection - induced fit* complex formation model.

Parameterizing these correlated motions, describing the corresponding energy landscapes, as well as handling collections of conformations pose challenging algorithmic problems.

At the side-chain level, the question of improving rotamer libraries is still of interest [23]. This question is essentially a clustering problem in the parameter space describing the side-chains conformations.

At the atomic level, flexibility is essentially investigated resorting to methods based on a classical potential energy (molecular dynamics), and (inverse) kinematics. A molecular dynamics simulation provides a point cloud sampling the conformational landscape of the molecular system investigated, as each step in the simulation corresponds to one point in the parameter space describing the system (the conformational space) [38]. The standard methodology to analyze such a point cloud consists of resorting to normal modes. Recently, though, more elaborate methods resorting to more local analysis [34], to Morse theory [29] and to analysis of meta-stable states of time series [30] have been proposed.

3.5. Algorithmic Foundations

Keywords: Computational geometry, computational topology, optimization, data analysis.

Making a stride towards a better understanding of the biophysical questions discussed in the previous sections requires various methodological developments, which we briefly discuss now.

3.5.1. Modeling Interfaces and Contacts

In modeling interfaces and contacts, one may favor geometric or topological information.

On the geometric side, the problem of modeling contacts at the atomic level is tantamount to encoding multi-body relations between an atom and its neighbors. On the one hand, one may use an encoding of neighborhoods based on geometric constructions such as Voronoi diagrams (affine or curved) or arrangements of balls. On the other hand, one may resort to clustering strategies in higher dimensional spaces, as the p neighbors of a given atom are represented by $3p - 6$ degrees of freedom – the neighborhood being invariant upon rigid motions. The information gathered while modeling contacts can further be integrated into interface models.

On the topological side, one may favor constructions which remain stable if each atom in a structure *retains* the same neighbors, even though the 3D positions of these neighbors change to some extent. This process is observed in flexible docking cases, and call for the development of methods to encode and compare shapes undergoing tame geometric deformations.

3.5.2. Modeling Macro-molecular Assemblies

In dealing with large assemblies, a number of methodological developments are called for.

[0] Assuming local forces are prominent, which in turn subsumes electrostatic interactions are not prominent.

On the experimental side, of particular interest is the disambiguation of proteomics signals. For example, TAP and mass spectrometry data call for the development of combinatorial algorithms aiming at unraveling pairwise contacts between proteins within an assembly. Likewise, density maps coming from electron microscopy, which are often of intermediate resolution (5-10Å) call the development of noise resilient segmentation and interpretation algorithms. The results produced by such algorithms can further be used to guide the docking of high resolutions crystal structures into maps.

As for modeling, two classes of developments are particularly stimulating. The first one is concerned with the design of algorithms performing reconstruction by data integration, a process reminiscent from non convex optimization. The second one encompasses assessment methods, in order to single out the reconstructions which best comply with the experimental data. For that endeavor, the development of geometric and topological models accommodating uncertainties is particularly important.

3.5.3. Modeling the Flexibility of Macro-molecules

Given a sampling on an energy landscape, a number of fundamental issues actually arise: how does the point cloud describe the topography of the energy landscape (a question reminiscent from Morse theory)? Can one infer the effective number of degrees of freedom of the system over the simulation, and is this number varying? Answers to these questions would be of major interest to refine our understanding of folding and docking, with applications to the prediction of structural properties. It should be noted in passing that such questions are probably related to modeling phase transitions in statistical physics where geometric and topological methods are being used [33].

From an algorithmic standpoint, such questions are reminiscent of *shape learning*. Given a collection of samples on an (unknown) *model*, *learning* consists of guessing the model from the samples – the result of this process may be called the *reconstruction*. In doing so, two types of guarantees are sought: topologically speaking, the reconstruction and the model should (ideally!) be isotopic; geometrically speaking, their Hausdorff distance should be small. Motivated by applications in Computer Aided Geometric Design, surface reconstruction triggered a major activity in the Computational Geometry community over the past ten years [5]. Aside from applications, reconstruction raises a number of deep issues: the study of distance functions to the model and to the samples, and their comparison; the study of Morse-like constructions stemming from distance functions to points; the analysis of topological invariants of the model and the samples, and their comparison.

AMIB Project-Team

3. Research Program

3.1. RNA

At the secondary structure level, we contributed novel generic techniques applicable to dynamic programming and statistical sampling, and applied them to design novel efficient algorithms for probing the conformational space. Another originality of our approach is that we cover a wide range of scales for RNA structure representation. For each scale (atomic, sequence, secondary and tertiary structure...) cutting-edge algorithmic strategies and accurate and efficient tools have been developed or are under development. This offers a new view on the complexity of RNA structure and function that will certainly provide valuable insights for biological studies.

3D modeling was supported by the Digiteo project JAPARIN-3D. Statistical potentials were supported by CARNAGE and ITSNAP.

3.1.1. Dynamic programming and complexity

Participants: Alain Denise, Yann Ponty, Antoine Soulé.

Common activity with J. Waldispühl (McGill).

Ever since the seminal work of Zuker and Stiegler, the field of RNA bioinformatics has been characterized by a strong emphasis on the secondary structure. This discrete abstraction of the 3D conformation of RNA has paved the way for a development of quantitative approaches in RNA computational biology, revealing unexpected connections between combinatorics and molecular biology. Using our strong background in enumerative combinatorics, we propose generic and efficient algorithms, both for sampling and counting structures using dynamic programming. These general techniques have been applied to study the sequence-structure relationship [77], the correction of pyrosequencing errors [71], and the efficient detection of multi-stable RNAs (riboswitches) [72], [73].

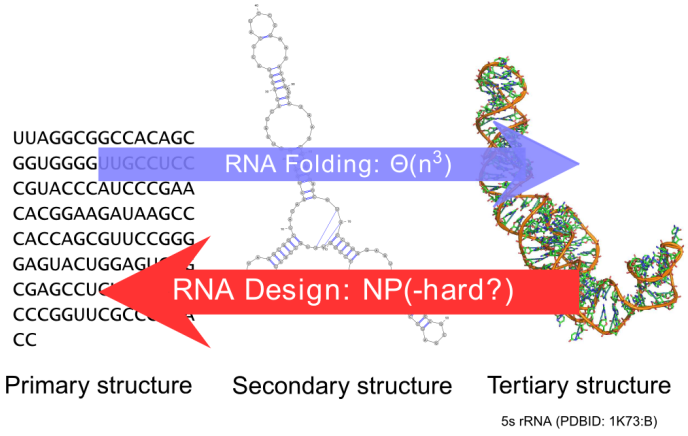


Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

3.1.2. RNA design.

Participants: Alain Denise, Vincent Le Gallic, Yann Ponty.

Joint project with S. Vialette (Marne-la-Vallée), J. Waldispühl (McGill) and Y. Zhang (Wuhan).

It is a natural pursue to build on our understanding of the secondary structure to construct artificial RNAs performing predetermined functions, ultimately targeting therapeutic and synthetic biology applications. Towards this goal, a key element is the design of RNA sequences that fold into a predetermined secondary structure, according to established energy models (inverse-folding problem). Quite surprisingly, and despite two decades of studies of the problem, the computational complexity of the inverse-folding problem is currently unknown.

Within our group, we offer a new methodology, based on weighted random generation [54] and multidimensional Boltzmann sampling, for this problem. Initially lifting the constraint of folding back into the target structure, we explored the random generation of sequences that are compatible with the target, using a probability distribution which favors exponentially sequences of high affinity towards the target. A simple posterior rejection step selects sequences that effectively fold back into the latter, resulting in a *global sampling* pipeline that showed comparable performances to its competitors based on local search [60].

3.1.3. Towards 3D modeling of large molecules

Participants: Alain Denise, Mélanie Boudard.

Joint project with D. Barth (Versailles) and J. Cohen (Paris-Sud).

The modeling of large RNA 3D structures, that is predicting the three-dimensional structure of a given RNA sequence, relies on two complementary approaches. The approach by homology is used when the structure of a sequence homologous to the sequence of interest has already been resolved experimentally. The main problem then is to calculate an alignment between the known structure and the sequence. The ab initio approach is required when no homologous structure is known for the sequence of interest (or for some parts of it). We work in both directions.

3.1.4. Statistical and robotics-inspired models for structure and dynamics

Participants: Julie Bernauer, Rasmus Fonseca.

Despite being able to correctly model small globular proteins, the computational structural biology community still craves for efficient force fields and scoring functions for prediction but also good sampling and dynamics strategies.

Our current and future efforts towards knowledge-based scoring function and ion location prediction have been described in 3.1.4 .

Over the last two decades a strong connection between robotics and computational structural biology has emerged, in which internal coordinates of proteins are interpreted as a kinematic linkage with rotatable bonds as joints and corresponding groups of atoms as links [76], [51], [64], [63]. Initially, fragments in proteins limited to tens of residues were modeled as a kinematic linkage, but this approach has been extended to encompass (multi-domain) proteins [62]. For RNA, progress in this direction has been realized as well. A kinematics-based conformational sampling algorithm, KGS, for loops was recently developed [58], but it does not fully utilize the potential of a kinematic model. It breaks and recloses loops using six torsional degrees of freedom, which results in a finite number of solutions. The discrete nature of the solution set in the conformational space makes difficult an optimization of a target function with a gradient descent method. Our methods overcome this limitation by performing a conformational sampling and optimization in a co-dimension 6 subspace. Fragments remain closed, but these methods are limited to proteins. Our objective is to extend the approach proposed in [58], [76] to nucleic acids and protein/nucleic acid complexes with a view towards improving structure determination of nucleic acids and their complexes and in silico docking experiments of protein/RNA complexes. For that purpose, we have developed a generic strategy for differentiable statistical potentials [2], [74] that can be directly integrated in the procedure.

Results from in silico docking experiments will also directly benefit structure determination of complexes which, in turn, will provide structural insights in nucleic acid and protein/nucleic acid complexes. From the small proof-of-concept single chain protein implementation of the KGS strategy, we have developed a robust preliminary implementation that can handle RNA and will be further developed to account for multi-chain molecules. Rasmus Fonseca, post-doctoral scholar in the project is currently performing an extensive computational and biological validation.

3.2. Sequences

Participants: Alain Denise, Mireille Régnier, Yann Ponty, Jean-Marc Steyaert, Alice Héliou, Daria Iakovishina, Antoine Soulé.

String searching and pattern matching is a classical area in computer science, enhanced by potential applications to genomic sequences. In CPM/SPIRE community, a focus is given to general string algorithms and associated data structures with their theoretical complexity. Our group specialized in a formalization based on languages, weighted by a probabilistic model. Team members have a common expertise in enumeration and random generation of combinatorial sequences or structures, that are *admissible* according to some given constraints. A special attention is paid to the actual computability of formula or the efficiency of structures design, possibly to be reused in external software.

As a whole, motif detection in genomic sequences is a hot subject in computational biology that allows to address some key questions such as chromosome dynamics or annotation. This area is being renewed by high throughput data and assembly issues. New constraints, such as energy conditions, or sequencing errors and amplification bias that are technology dependent, must be introduced in the models. An other aim is to combine statistical sampling with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [66]. In general, in the future, our methods for sampling and sequence data analysis should be extended to take into account such constraints, that are continuously evolving.

3.2.1. Combinatorics of motifs

Participants: Mireille Régnier, Alice Héliou, Daria Iakovishina.

Besides applications [5] of analytic combinatorics to computational biology problems, the team addressed general combinatorial problems on words and fundamental issues on languages and data structures.

Molecular interactions often involve specific motifs. One may cite protein-DNA (cis-regulation), protein-protein (docking), RNA-RNA (miRNA, frameshift, circularisation). Motif detection combines an algorithmic search of potential sites and a significance assessment. Assessment significance requires a quantitative criterium. It is generally accepted that the p-value is a reliable tool that outperforms older criteria such as the z-score. AMIB develops a long term research on word combinatorics. In the recent years, a general scheme of derivation of analytic formula for the pvalue under different constraints (k-occurrence, first occurrence, overrepresentation in large sequences,...) has been provided. It relies on a representation of word overlaps in a graph [40]. Recursive equations to compute pvalues may be reduced to a traversal of that graph, leading to a linear algorithm. It allows for a derivation of pvalues, decreasing the space and time complexity of the generating function approach or previous probabilistic weighted automata.

In the mean time, continuous sequences of overlapping words, currently named *clumps* or *clusters* turn out to be crucial in random words counting. Notably, they play a fundamental role in the Chen-Stein method of compound Poisson approximation. A first characterization was proposed by Nicodème and al. and this work is currently extended.

This research area is widened by new problems arising from *de novo* genome assembly or re-assembly. For example, unique mappability of short reads strongly depends of the repetition of words. Although the average values for the length have been studied for long under different constraints, their distribution or profile remained unknown until the seminal paper [67] which provides formulae for binary tries. A collaboration has been started with LOB at Ecole Polytechnique to check these formulae on real data, namely Archae genomes (internship of D. Busatto-Gaston). This collaboration has been extended since LOB bought a sequencing machineand a co-advisd thesis (Alice Héliou) on circular RNA characterization hasjust started.

As a third example, one objective is to develop a model of errors, including a statistical model, that takes into account the quality of data for the different sequencing technologies, and their volume. This is the subject of an international collaboration with V. Makeev's lab (IoGene, Moscow) and MAGNOME project-team. Finally, Next Generation Sequencing open the way to the study of structural variants in the genome, as recently described in [48]. Defining a probabilistic model that takes into account main dependencies -such as the GC content- is a task of D. Iakovishina's thesis, to be defended in 2015, in a collaboration with V. Boeva (Curie Institute).

3.2.2. Random generation

Participants: Alain Denise, Yann Ponty.

Analytical methods may fail when both sequential and structural constraints of sequences are to be modelled or, more generally, when molecular *structures* such as RNA structures have to be handled. The random generation of combinatorial objects is a natural, alternative, framework to assess the significance of observed phenomena. General and efficient techniques have been developed over the last decades to draw objects uniformly at random from an abstract specification. However, in the context of biological sequences and structures, the uniformity assumption becomes unrealistic, and one has to consider non-uniform distributions in order to derive relevant estimates. Typically, context-free grammars can handle certain kinds of long-range interactions such as base pairings in secondary RNA structures.

In 2005, a new paradigm appeared in the *ab initio* secondary structure prediction [55]: instead of formulating the problem as a classic optimization, this new approach uses statistical sampling within the space of solutions. Besides giving better, more robust, results, it allows for a fruitful adaptation of tools and algorithms derived in a purely combinatorial setting. Indeed, we have done significant and original progress in this area recently [68], [5], including combinatorial models for structures with pseudoknots. Our aim is to combine this paradigm with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [66].

Besides, our work on random generation is also applied in a different fields, namely software testing and model-checking, in a continuing collaboration with the Fortesse group at LRI [53], [65].

3.3. Geometry and machine learning for 3D interaction prediction

Participants: Julie Bernauer, Jean-Marc Steyaert, Christine Froidevaux, Adrien Guilhot-Gaudeffroy, Amélie Héliou.

The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. This is specially challenging as structure flexibility is key and multi-scale modelling [47], [57] and efficient code are essential [61].

Our project covers various aspects of biological macromolecule structure and interaction modelling and analysis. First protein structure prediction is addressed through combinatorics. The dynamics of these types of structures is also studied using statistical and robotics inspired strategies. Both provide a good starting point to perform 3D interaction modelling, accurate structure and dynamics being essential. Modelling is then raised to the cell level by studying large protein interaction networks and also the dynamics of molecular pathways.

Our group benefits from a good collaboration network, mainly at Stanford University (USA), HKUST (Hong-Kong) and McGill (Canada). The computational expertise in this field of computational structural biology is represented in a few large groups in the world (e.g. Pande lab at Stanford, Baker lab at U.Washington) that have both dry and wet labs. We also contributed to the CAPRI experiment organized by leading member of an international community we have been involved in for some time [56]. At Inria, our interest for structural biology is shared by the ABS project-team. A work by D. Ritchie in the ORPAILLEUR project-team (see [44] led to a joint publication with T. Bourquard and J. Azé. Our activities are however now more centered around protein-nucleic acid interactions, multi-scale analysis, robotics inspired strategies and machine learning than protein-protein interactions, algorithms and geometry. We also shared a common interest for large biomolecules and their dynamics with the NANO-D project team and their adaptative sampling strategy. As

a whole, we contribute to the development of geometric and machine learning strategies for macromolecular docking.

3.3.1. Combinatorial models for the structure of proteins

Protein structure prediction has been and still is extensively studied. Computational approaches have shown interesting results for globular proteins but transmembrane proteins remain a difficult case.

Transmembrane beta-barrel proteins (TMB) account for 20 to 30% of identified proteins in a genome but, due to difficulties with standard experimental techniques, they are only 2% of the RCSB Protein Data Bank. As TMB perform many vital functions, the prediction of their structure is a challenge for life sciences, while the small number of known structures prohibits knowledge-based methods for structure prediction.

As barrel proteins are strongly structured objects, model based methodologies are an interesting alternative to these conventional methods. Jérome Waldisphül's thesis at LIX had opened this track for the common case where a protein folds respecting the order of the sequence, leaving a structure where each strand is bound to the preceding and succeeding ones. The matching constraints were expressed by a grammatical model, for which relatively simple dynamic programming schemes exist.

However, more sophisticated schemes are required when the arrangements of the strands along the barrel do not follow their order in the sequence, as it is the case for *Greek key* or *Jelly roll* motifs. The prediction algorithm may then be driven by a permutation on the order of the bonded strands. In his thesis [75], Van Du Tran developed a methodology for compiling a given permutation into a dynamic programming scheme that may predict the folding of sequences into the corresponding TMB secondary structure. Polynomial complexity upper bounds follow from the calculated DP scheme. Through tree decompositions of the graph that expresses constraints between strands in the barrel, better schemes were investigated in [75].

The efficiently obtained 3D structures provide a good model for further 3D and interaction analyses.

3.3.2. 3D interaction prediction

To better model complexes, various aspects of the scoring problem for protein-protein docking need being addressed [56]. It is also of great interest to introduce a hierarchical analysis of the original complex three-dimensional structures used for learning, obtained by clustering.

A protein-protein docking procedure traditionally consists in two successive tasks: a search algorithm generates a large number of candidate solutions, and then a scoring function is used to rank them in order to extract a native-like conformation. We demonstrated that, using Voronoi constructions and a defined set of parameters, we could optimize an accurate scoring function and interaction detection [46]. We also focused on developing other geometric constructions for that purpose: being related to the Voronoi construction, the Laguerre tessellation was expected to better represent the physico-chemical properties of the partners. It also allows a fast computation without losing the intrinsic properties of the biological objects. In [49], we compare both constructions. We also worked on introducing a hierarchical analysis of the original complex three-dimensional structures used for learning, obtained by clustering. Using this clustering model, in combination with a strong emphasis on the design of efficient complex filters collaborative filtering, we can optimize the scoring functions and get more accurate solutions [50].

These techniques have been extended to the analysis of protein-nucleic acid complexes : developments and tests are performed by A. Guilhot (See figure 2) in his PhD thesis.

3.4. Data Integration

Participants: Christine Froidevaux, Alain Denise, Sarah Cohen-Boulakia, Bryan Brancotte, Jiuqiang Chen.

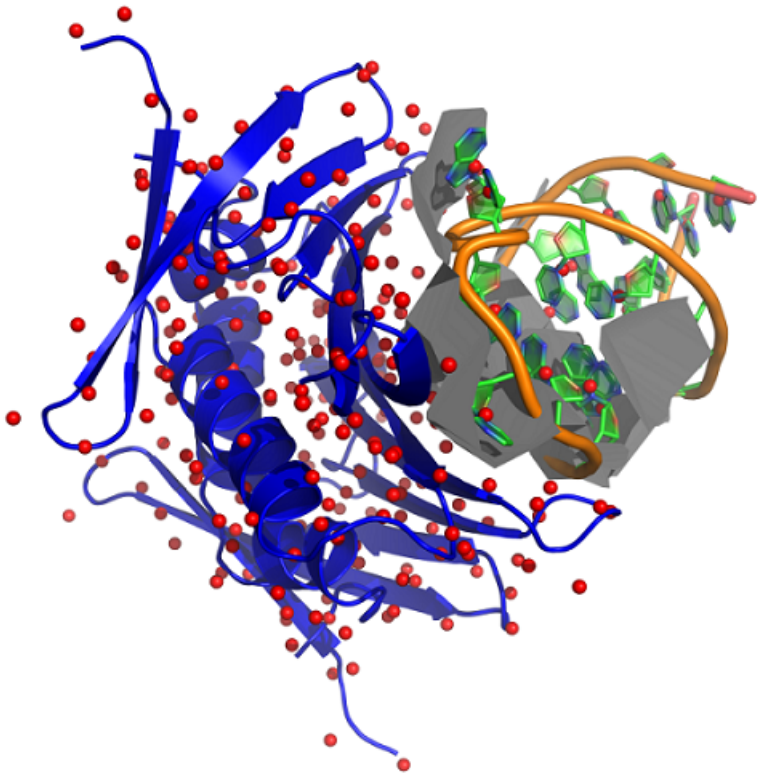


Figure 2. Coarse-grained representation and Voronoi interface model of a PP7 coat protein bound to an RNA hairpin (PDB code 2qux). The Voronoi model captures the features of the interactions such as stacking, even at the coarse-grained level.

Faced with the inherent features of biological and biomedical data, researchers from the database and artificial intelligence communities have joined together to form a community dedicated to the study of the specific problems posed by integrating life sciences data. With the deluge of new sequenced genome sequences and the amount of data produced by high-throughput approaches, the need to cross and compare massive and heterogeneous data is more important than ever to improve functional annotation and design biological networks. Challenges are numerous. One may cite the need to provide support to scientists to perform and share complex and reproducible complex biological analyses. A special attention is paid to the more specific domain of scientific workflows management and ranking biological data. One aims at exploring the relationships between those two domains, from the investigation of various specific problems posed by ranking scientific workflows to the problem of considering consensus workflows.

3.4.1. Designing and Comparing Scientific workflows

Participants: Sarah Cohen-Boulakia, Christine Froidevaux, Jiuqiang Chen.

Scientific workflows management systems are increasingly used to specify and manage bioinformatics experiments. Their programming model appeals to bioinformaticians, who use them to easily specify complex data processing pipelines. Such a model is underpinned by a graph structure, where nodes represent bioinformatics tasks and links represent the dataflow. As underlined both in a study and a review of existing approaches, the complexity of such graph structures is increasing over time, making them more difficult to share and reuse.

One of the major current challenges is thus to provide means to reduce the structural complexity of workflows while ensuring that any structural transformation will not have any impact on the executions of the transformed workflows, that is, preserving *provenance*.

3.4.2. Ranking biological data

Participants: Alain Denise, Sarah Cohen-Boulakia, Bryan Brancotte.

We are addressing the increase of the number of resources available. The BIOGUIDE project aim at helping user navigation in the maze of available biological sources. More recently, a second problem was tackled: the number of answers returned by even one single queried biological resource may be too large for the user to deal with. We have provided solutions for ranking biological data. The main difficulty lies in considering various ranking criteria (recent data first, popular data first, curated data first...). Many approaches combine ranking criteria to design a ranking function, possibly leading to arbitrary choices made in the way of combining the ranking criteria. Instead, in collaboration with the University of Montreal, we have proposed to follow a *median ranking approach* named BIOCONSERT (for generating Biological Consensus ranking with ties): considering as many rankings as they are ranking criteria for the same data set, and providing a consensus ranking that minimizes the disagreements between the input rankings. We have shown the benefit of using median ranking in several biological settings.

Additionally, in a close collaboration with the Institut Curie, we have also developed the GENEVALORIZATION tool that ranks a list of genes of interest given as input with respect to a set of keywords representing the context of study. Here the single ranking criterion considered for each gene is the number of publications in PubMed co-citing the gene name and the keywords. The tool is able to make use of the MeSH taxonomy when considering the keywords and the dictionary of gene names and aliases for the gene names.

3.5. Systems Biology

Participants: Patrick Amar, Sarah Cohen-Boulakia, Alain Denise, Christine Froidevaux, Loic Paulevé, Sabine Pérès, Jean-Marc Steyaert, Erwan Bigan, Adrien Rougny.

Systems Biology involves the systematic study of complex interactions in biological systems using an integrative approach. The goal is to find new emergent properties that may arise from the systemic view in order to understand the wide variety of processes that happen in a biological system. Systems Biology activity can be seen as a cycle composed of theory, computational modelling to propose a hypothesis about a biological process, experimental validation, and use of the experimental results to refine or invalidate the computational model (or even the whole theory). During the past five years, new questions and research domains have been identified, and some members of the team have reoriented a part of their activities on these questions.

Three main types of problems have been studied: metabolic networks, signaling networks and more recently synthetic biology. Networks - have become popular since many crucial problems, coming from biology, medecine, pharmacology, are nowadays stated in these terms: a great number of them are issued from the cancer phenomenom and the will to enhance our understanding in order to propose more efficient therapeutic issues. Metabolism has received the major attention since it concerns a large variety of topics and several methods that have been proposed. Depending on the nature of the biological problem, several methods can be used : discrete deterministic, stochastic, combinatorial, up to continuous differential. Also, the recent rise of synthetic biology proposes similar challenges aiming at improving the production of energy by means of biological systems or at getting more efficient medicamental treatments, for instance.

3.5.1. Topological analysis of metabolic networks

Participant: Sabine Pérès.

Elementary flux mode analysis is a powerful tool for the theoretical study of simple metabolic networks. However, when the networks are complex, the determination of elementary flux modes leads to a combinatorial explosion of their number which prevents from drawing simple conclusions from their analysis. Since the concept of elementary flux mode analysis was introduced in 1994, there has been an important and ongoing effort to develop more efficient algorithms. However, these methods share a common bottleneck: they enumerate all the elementary flux modes which make the computation impossible when the metabolic network is large and only few works try to search only elementary flux mode with specific properties. We have shown that enumerating all the elementary flux modes is not necessary in many cases and it is possible to directly query the network instead with an appropriate tool. For ensuring a good query time, we have relied on a state of the art SAT solver, working on a propositional encoding of elementary flux mode, and enriched with a simple SMT-like solver ensuring elementary flux mode consistency with stoichiometric constraints. We have illustrated our new framework by providing experimental evidences of almost immediate answer times on a non trivial metabolic network [45], [70].

3.5.2. Signaling networks

Participants: Sarah Cohen-Boulakia, Christine Froidevaux, Adrien Rougny.

Signaling pathways involving G protein-coupled receptors (GPCR) are excellent targets in pharmacogenomics research. Large amounts of experiments are available in this context while globally interpreting all the experimental data remains a very challenging task for biologists. Our goal is to help the understanding of signaling pathways involving (GPCR) and to provide means to semi-automatically construct the signaling networks.

We have introduced a logic-based method to infer molecular networks and show how it allows inferring signaling networks from the design of a knowledge base. Provenance of inferred data has been carefully collected, allowing quality evaluation. Our method (i) takes into account various kinds of biological experiments and their origin; (ii) mimics the scientist's reasoning within a first-order logic setting; (iii) specifies precisely the kind of interaction between the molecules; (iv) provides the user with the provenance of each interaction; (v) automatically builds and draws the inferred network [43].

Observe that a logic-based formalisation is used as in some works carried out in INRIA team DYLISS. AMIB aim is different, as the design of the network lies on a knowledge-based system describing experimental facts and ontological relationships on backgound knowledge, together with a set of generic and expressive rules, that mimick the expert's reasoning.

This is a collaboration with A. Poupon (INRA-BIOS, Tours) that was supported by an INRA-INRIA starting grant in 2011-2012.

3.5.3. Modelling and Simulation

Participants: Patrick Amar, Sarah Cohen-Boulakia, Loic Paulevé, Jean-Marc Steyaert, Erwan Bigan.

A great number of methods have been proposed for the study of the behavior of large biological systems. The first one is based on a discrete and direct simulation of the various interactions between the reactants using an entity-centered approach; the second one implements a very efficient variant of the Gillespie stochastic algorithm that can be mixed with the entity-centered method to get the best of both worlds; the third one uses differential equations automatically generated from the set of reactions defining the network.

These three methods have been implemented in an integrated tool, the HSIM system [41]. It mimics the interactions of biomolecules in an environment modelling the membranes and compartments found in real cells. It has been applied to the modelling of the circadian clock of the cyanobacterium, and we have shown pertinent results regarding the spontaneous appearance of oscillations and the factors governing their period [42].

3.5.3.1. Synthetic biology

Synthetic biology begins to be a very popular domain of research. Genetic engineering is a good example of synthetic biology, organisms are artificially modified to boost the production of compounds that might be used in the medical or industrial domains. We have been focused on using synthetic biology for medical diagnostic purposes. In a collaboration with the SYSDIAGLab (UMR 3145) at Montpellier, P. Amar participates at the COMPUBIOTIC project. The goal is to design, test and build an artificial embedded biological nano-computer in order to detect the biological markers of some human pathologies (colorectal cancer, diabetic nephropathy, etc.). This nano-computer is a small vesicle containing specific enzymes and membrane receptors. These components are chosen in a way that their interactions can sense and report the presence in the environment of molecules involved in the human pathologies tageted. We plan to design a dedicated software suite to help the design and validation of this artificial nano-computer. HSIM is used to help the design and to test qualitatively and quantitatively this *"biological computer"* before *in vitro*.

3.5.3.2. Evaluating metabolic networks

It is now well established in the medical world that the metabolism of organs depends crucially of the way the calls consume oxygene, glucose and the various metabolites that allow them to grow and duplicate. A particular variety of cells, tumour cells, is of major interest. In collaboration with L. Schwartz (AP-HP) and biologists from INSERM-INRA Clermont-Theix we have started a project aiming at identifying the important points in the metabolic machinery that command the changes in behaviour. The main difficulties come from the fact that biologists have listed dozens of concurrent cycles that can be activated alternatively or simultaneously, and that the dynamic characteristics of the chemical reactions are not known accurately.

Given the set of biochemical reactions that describe a metabolic function (e.g. glycolysis, phospholipids' synthesis, etc.) we translate them into a set of o.d.e's whose general form is most often of the Michaelis-Menten type but whose coefficients are usuall very badly determined. The challenge is therefore to extract information as to the system's behavior while making reasonable asumptions on the ranges of values of the parameters. It is sometimes possible to prove mathematically the global stability, but it is also possible to establish it locally in large subdomains by means of simulations.Our program `Mpas` (Metabolic Pathway Analyser Software) renders the translation in terms of a systems of o.d.e's automatic, leading to easy, almost automatic simulations. Furthermore we have developed a method of systematic analysis of the systems in order to characterize those reactants which determine the possible behaviors: usually they are enzymes whose high or low concentrations force the activation of one of the possible branches of the metabolic pathways. A first set of situations has been validated with a research INSERM-INRA team based in Clermont-Ferrand. In her PhD thesis, defended in 2011, M. Behzadi proved mathematically the decisive influence of the enzyme PEMT on the Choline/Ethylamine cycles.

3.5.3.3. Comparison of Metabolic Networks

We study the interest of *fungi* for biomass transformation. Cellulose, hemicellulose and lignin are the main components of plant biomass. Their transformation represent a key energy challenges of the 21st century and should eventually allow the production of high value new compounds, such as wood or liquid biofuels (gas or bioethanol). Among the boring organisms, two groups of fungi differ in how they destroy the wood compounds. Analysing new fungi genomes can allow the discover of new species of high interest for

bio-transformation. For a better understanding of how the fungal enzymes facilitates degradation of plant biomass, we conduct a large-scale analysis of the metabolism of fungi. Machine learning approaches such like hierarchical rules prediction are being studied to find new enzymes allowing the transformation of biomass. The KEGG database http://www.genome.jp/kegg/ contains pathways related to fungi and other species. By analysing these known pathways with rules mining approaches, we aim to predict new enzymes activities.

ANGE Project-Team

3. Research Program

3.1. Overview

The research activities carried out within the ANGE team strongly couple the development of methodological tools with applications to real–life problems and the transfer of numerical codes. The main purpose is to obtain new models adapted to the physical phenomena at stake, identify the main properties that reflect the physical sense of the models (uniqueness, conservativity, entropy dissipation, ...) and propose effective numerical methods to estimate their solution in complex configurations (multi-dimensional, unstructured meshes, well-balanced, ...).

The difficulties arising in gravity driven flow studies are threefold.

- Models and equations encountered in fluid mechanics (typically the free surface Navier-Stokes equations) are complex to analyze and solve.
- The underlying phenomena often take place over large domains with very heterogeneous length scales (size of the domain, mean depth, wave length,...) and distinct time scales, *e.g.* coastal erosion, propagation of a tsunami,...
- These problems are multi-physics with strong couplings and nonlinearities.

3.2. Modelling and analysis

Hazardous flows are complex physical phenomena that can hardly be represented by shallow water type systems of partial differential equations (PDEs). In this domain, the research program is devoted to the derivation and analysis of reduced complexity models compared to the Navier-Stokes equations, but relaxing the shallow water assumptions. The main purpose is then to obtain models well-adapted to the physical phenomena at stake.

Even if the resulting models do not strictly belong to the family of hyperbolic systems, they exhibit hyperbolic features: the analysis and discretization techniques we intend to develop have connections with those used for hyperbolic conservation laws. It is worth noticing that the need for robust and efficient numerical procedures is reinforced by the smallness of dissipative effects in geophysical models which therefore generate singular solutions and instabilities.

On the one hand, the derivation of the Saint-Venant system from the Navier-Stokes equations is based on two approximations, so-called shallow water assumptions, namely

- the horizontal fluid velocity is well approximated by its mean value along the vertical direction,
- the pressure is hydrostatic or equivalently the vertical acceleration of the fluid can be neglected compared to the gravitational effects.

As a consequence the objective is to get rid of these two assumptions, one after the other, in order to obtain models accurately approximating the incompressible Euler or Navier-Stokes equations.

On the other hand, many applications require the coupling with non-hydrodynamic equations, as in the case of micro-algae production or erosion processes. These new equations comprise non-hyperbolic features and must rely on a special analysis.

3.2.1. Multilayer approach

As for the first shallow water assumption, *multi-layer* systems were proposed describing the flow as a superposition of Saint-Venant type systems [21], [24], [25]. Even if this approach has provided interesting results, layers are considered separate and non-miscible fluids, which imply strong limitation. That is why we proposed a slightly different approach [22], [23] based on Galerkin type decomposition along the vertical axis of all variables and leading, both for the model and its discretization, to more accurate results.

A kinetic representation of our multilayer model allows to derive robust numerical schemes endowed with properties such as: consistency, conservativity, positivity, preservation of equilibria,... It is one of the major achievements of the team but it needs to be analyzed and extended in several directions namely:

- The convergence of the multilayer system towards the hydrostatic Euler system as the number of layers goes to infinity is a critical point. It is not fully satisfactory to have only formal estimates of the convergence and sharp estimates would enable to guess the optimal number of layers.
- The introduction of several source terms due for instance to Coriolis forces or extra terms from changes of coordinates seems necessary. Their inclusion should lead to substantial modifications of the numerical scheme.
- Its hyperbolicity has not yet been proved and conversely the possible loss of hyperbolicity cannot be characterized. Similarly, the hyperbolic feature is essential in the propagation and generation of waves.

3.2.2. Non-hydrostatic models

The hydrostatic assumption consists in neglecting the vertical acceleration of the fluid. It is considered valid for a large class of geophysical flows but is restrictive in various situations where the dispersive effects (like wave propagation) cannot be neglected. For instance, when a wave reaches the coast, bathymetry variations give a vertical acceleration to the fluid that strongly modifies the wave characteristics and especially its height.

When processing an asymptotic expansion (w.r.t. the aspect ratio for shallow water flows) into the Navier-Stokes equations, we obtain at the leading order the Saint-Venant system. Going one step further leads to a vertically averaged version of the Euler/Navier-Stokes equations integrating the non-hydrostatic terms. This model has several advantages:

- it admits an energy balance law (that is not the case for most dispersive models available in the literature),
- it reduces to the Saint-Venant system when the non-hydrostatic pressure term vanishes,
- it consists in a set of conservation laws with source terms,
- it does not contain high order derivatives.

3.2.3. Multi-physics modelling

The coupling of hydrodynamic equations with other equations in order to model interactions between complex systems represents an important part of the team research. More precisely, three multi-physic systems are investigated. More details about the industrial impact of these studies are presented in the following section.

- To estimate the risk for infrastructures in coastal zone or close to a river, the resolution of the shallow water equations with moving bathymetry is necessary. The first step consisted in the study of an equation largely used in engineering science: The Exner equation. The analysis enabled to exhibit drawbacks of the coupled model such as the lack of energy conservation or the strong variations of the solution from small perturbations. A new formulation is proposed to avoid these drawbacks. The new model consists in a coupling between conservation laws and an elliptic equation, like the system Euler/Poisson, suggesting to use well-known strategies for the analysis and the numerical resolution. In addition, the new formulation is derived from classical complex rheology models and allowed physical phenomena such as threshold laws.
- Interaction between flows and floating structures is the challenge at the scale of the shallow water equations. This study needs a better understanding of the energy exchanges between the flow and the structure. The mathematical model of floating structures is very hard to solve numerically due to the non-penetration condition at the interface between the flow and the structure. It leads to infinite potential wave speeds that could not be solved with classical free surface numerical scheme. A relaxation model was derived to overcome this difficulty. It represents the interaction with the floating structure with a free surface model-type.

- If the interactions between hydrodynamics and biology phenomena are known through laboratory experiments, it is more difficult to predict the evolution, especially for the biological quantities, in a real and heterogeneous system. The objective is to model and reproduce the hydrodynamics modifications due to forcing term variations (in time and space). We are typically interested in phenomena such as eutrophication, development of harmful bacteria (cyanobacteria) and upwelling phenomena.

3.3. Numerical analysis

3.3.1. Non-hydrostatic scheme

The main challenge in the study of the non-hydrostatic model is to design a robust and efficient numerical scheme endowed with properties such as: positivity, wet/dry interfaces treatment, consistency. It has to be noticed that even if the non-hydrostatic model looks like an extension of the Saint-Venant system, most of the known techniques used in the hydrostatic case are not efficient as we recover strong difficulties encountered in incompressible fluid mechanics due to the extra pressure term. These difficulties are reinforced by the absence of viscous/dissipative terms.

3.3.2. Space decomposition and adaptive scheme

In the quest for a better balance between accuracy and efficiency, a strategy consists in the adaptation of models. Indeed, the systems of partial differential equations we consider result from a hierarchy of simplifying assumptions. However, some of these hypotheses may turn out to be irrelevant locally. The adaptation of models thus consists in determining areas where a simplified model (*e.g.* shallow water type) is valid and where it is not. In the latter case, we may go back to the "parent" model (*e.g.* Euler) in the corresponding area. This implies to know how to handle the coupling between the aforementioned models from both theoretical and numerical points of view. In particular, the numerical treatment of transmission conditions is a key point. It requires the estimation of characteristic values (Riemann invariant) which have to be determined according to the regime (torrential or fluvial).

3.3.3. Asymptotic-Preserving scheme for source terms

The hydrodynamic models comprise advection and sources terms. The conservation of the balance between the source terms, typically viscosity and friction, has a significant impact since the overall flow is generally a perturbation around one equilibrium. The design of numerical schemes able to preserve such balances is a challenge from both theoretical and industrial points of view.The concept of Asymptotic-Preserving (AP) methods is of great interest in order to overcome these issues.

Another difficulty occurs when a term, typically related to the pressure, becomes very large compared to the order of magnitude of the velocity. At this regime, namely the so-called *low Froude* (shallow water) or *low Mach* (Euler) regimes, the difference between the speed of the potential waves and the physical velocity makes classical numerical schemes not efficient: firstly because of the error of truncation which is inversely proportional to the small parameters, secondly because of the time step governed by the largest speed of the potential wave. AP methods made a breakthrough in the numerical resolution of asymptotic perturbations of partial-differential equations concerning the first point. The second one can be fixed using partially implicit scheme.

3.3.4. Multi-physics models

Coupling problems also arise within the fluid when it contains pollutants, density variations or biological species. For most situations, the interactions are small enough to use a splitting strategy and the classical numerical scheme for each sub-model, whether it be hydrodynamic or non-hydrodynamic.

The sediment transport raises interesting issues from a numerical aspect. This is an example of coupling between the flow and another phenomenon, namely the deformation of the bottom of the basin that can be carried out either by bed load where the sediment has its own velocity or suspended load in which the particles are mostly driven by the flow. This phenomenon involves different time scales and nonlinear retroactions; hence the need for accurate mechanical models and very robust numerical methods. In collaboration with industrial partners (EDF–LNHE), the team already works on the improvement of numerical methods for existing (mostly empirical) models but our aim is also to propose new (quite) simple models that contain important features and satisfy some basic mechanical requirements. The extension of our 3D models to the transport of weighted particles can also be here of great interest.

3.3.5. Data assimilation

Data assimilation consists in a coupling between a model and observation measurements. Developing robust data assimilation methods for hyperbolic-type conservation laws is a challenging subject. These PDEs indeed show no dissipation effects and the input of additional information in the model equations may introduce errors that propagate and create shocks. We have recently proposed a new approach based on the kinetic description of the conservation law. Hence, data assimilation is carried out at the kinetic level, using a Luenberger observer. Assimilation then resumes to the handling of a BGK type equation. The advantage of this framework is that we deal with a single "linear" equation instead of a nonlinear system and it is easy to recover the macroscopic variables. We are able to prove the convergence of the model towards the data in case of complete observations in space and time.

This work is done in collaboration with the M3DISIM Inria project-team. M. Doumic and B. Perthame (MAMBA) also participate.

ARAMIS Project-Team

3. Research Program

3.1. General aim

The overall aim of our project is to design new computational and mathematical approaches for studying brain structure (based on anatomical and diffusion MRI) and functional connectivity (based on EEG, MEG and intracerebral recordings). The goal is to transform raw unstructured images and signals into formalized, operational models such as geometric models of brain structures, statistical population models, and graph-theoretic models of brain connectivity. This general endeavor is addressed within the three following main objectives.

3.2. Modeling brain structure: from imaging to geometric models

Structural MRI (anatomical or diffusion-weighted) allows studying in vivo the anatomical architecture of the brain. Thanks to the constant advance of these imaging techniques, it is now possible to visualize various anatomical structures and lesions with a high spatial resolution. Computational neuroanatomy aims at building models of the structure of the human brain, based on MRI data. This general endeavor requires addressing the following methodological issues: i) the extraction of geometrical objects (anatomical structures, lesions, white matter tracks...) from anatomical and diffusion-weighted MRI; ii) the design of a coherent mathematical framework to model anatomical shapes and compare them across individuals. Within this context, we pursue the following objectives.

First, we aim to develop new methods to segment anatomical structures and lesions. We are most specifically interested in the hippocampus, a structure playing a crucial role in Alzheimer's disease, and in lesions of vascular origin (such as white matter hyperintensities and microbleeds). We pay particular attention to the robustness of the approaches with respect to normal and pathological anatomical variability and with respect to differences in acquisition protocols, for application to multicenter studies. We dedicate specific efforts to the validation on large populations of coming from patients data acquired in multiple centers.

Then, we develop approaches to estimate templates from populations and compare anatomical shapes, based on a diffeomorphic deformation framework and matching of distributions. These methods allow the estimation of a prototype configuration (called template) that is representative of a collection of anatomical data. The matching of this template to each observation gives a characterization of the anatomical variability within the population, which is used to define statistics. In particular, we aim to design approaches that can integrate multiple objects and modalities, across different spatial scales.

3.3. Modeling dynamical brain networks

Functional imaging techniques (EEG, MEG and fMRI) allow characterizing the statistical interactions between the activities of different brain areas, i.e. functional connectivity. Functional integration of spatially distributed brain regions is a well-known mechanism underlying various cognitive and perceptual tasks. Indeed, mounting evidence suggests that impairment of such mechanisms might be the first step of a chain of events triggering several neurological disorders, such as the abnormal synchronization of epileptic activities. Naturally, neuroimaging studies investigating functional connectivity in the brain have become increasingly prevalent.

Our team develops a framework for the characterization of brain connectivity patterns, based on connectivity descriptors from the theory of complex networks. The description of the connectivity structure of neural networks is able to characterize for instance, the configuration of links associated with rapid/abnormal synchronization and information transfer, wiring costs, resilience to certain types of damage, as well as the balance between local processing and global integration. Furthermore, we propose to extend this framework to study the reconfiguration of networks over time. Indeed, neurophysiological data are often gathered from longitudinal recording sessions of the same subject to study the adaptive reconfiguration of brain connectivity. Finally, connectivity networks are usually extracted from different brain imaging modalities (MEG, EEG, fMRI or DTI) separately. Methods for combining the information carried by these different networks are still missing. We thus propose to combine connectivity patterns extracted from each modality for a more comprehensive characterization of networks.

3.4. Methodologies for large-scale datasets

Until recently, neuroimaging studies were often restricted to series of about 20-30 patients. As a result, such studies had a limited statistical power and could not adequately model the variability of populations. Thanks to wider accessibility of neuroimaging devices and important public and private funding, large-scale studies including several hundreds of patients have emerged in the past years. In the field of Alzheimer's disease (AD) for instance, one can cite the Alzheimer's Disease Neuroimaging Initiative (ADNI) including about 800 subjects (patients with AD or mild cognitive impairment (MCI) and healthy controls) or the French cohort MEMENTO including about 2000 subjects with memory complaint. These are most often multicenter studies in which patients are recruited over different centers and images acquired on different scanners. Moreover, cohort studies include a longitudinal component: for each subject, multiple images are acquired at different time points. Finally, such datasets often include multimodal data: neuroimaging, clinical data, cognitive tests and genomics data. These datasets are complex, high-dimensional and often heterogeneous, and thus require the development of new methodologies to be fully exploited.

In this context, our objectives are:

- to develop methodologies to acquire and standardize multicenter neuroimaging data;
- to develop imaging biomarkers based on machine learning and longitudinal models;
- to design multimodal analysis approaches for bridging anatomical models and genomics.

The first two aspects focus on neuroimaging and are tightly linked with the CATI project. The last one builds on our previous expertise in morphometry and machine learning, but aims at opening new research avenues combining imaging and "omics" data. This is developed in strong collaboration with the new biostatistics/bioinformatics platform of the IHU-A-ICM.

ASCLEPIOS Project-Team

3. Research Program

3.1. Introduction

Tremendous progress has been made in the automated analysis of biomedical images during the past two decades [56]. Readers who are neophytes to the field of medical imaging will find an interesting presentation of acquisition techniques of the main medical imaging modalities in [48], [46]. Regarding target applications, a good review of the state of the art can be found in the book *Computer Integrated Surgery* [44], in N. Ayache's article [51] and in the more recent syntheses [52] [56]. The scientific journals *Medical Image Analysis* [39], *Transactions on Medical Imaging* [45], and *Computer Assisted Surgery* [47] are also good reference material. One can have a good vision of the state of the art with the proceedings of the most recent conferences MICCAI'2010 (Medical Image Computing and Computer Assisted Intervention) [42], [43] or ISBI'2010 (Int. Symp. on Biomedical Imaging) [41].

For instance, for rigid parts of the body like the head, it is now possible to fuse in a completely automated manner images of the same patient taken from different imaging modalities (e.g. anatomical and functional), or to track the evolution of a pathology through the automated registration and comparison of a series of images taken at distant time instants [57], [67]. It is also possible to obtain from a Magnetic Resonance Image (MRI) of the head a reasonable segmentation into skull tissues, white matter, grey matter, and cerebro-spinal fluid [70], or to measure some functional properties of the heart from dynamic sequences of Magnetic Resonance [50], Ultrasound or Nuclear Medicine images [58].

Despite these advances and successes, statistical models of anatomy are still very crude, resulting in poor registration results in deformable regions of the body, or between different subjects. If some algorithms exploit physical modeling of the image acquisition process, only a few actually model the physical or even physiological properties of the human body itself. Coupling biomedical image analysis with anatomical and physiological models of the human body could not only provide a better comprehension of observed images and signals, but also more efficient tools for detecting anomalies, predicting evolutions, simulating and assessing therapies.

3.2. Medical Image Analysis

The quality of biomedical images tends to improve constantly (better spatial and temporal resolution, better signal to noise ratio). Not only are the images multidimensional (3 spatial coordinates and possibly one temporal dimension), but medical protocols tend to include multi-sequence (or multi-parametric) [0] and multi-modal images [0] for each single patient.

[0]Multisequence (or multiparametric) imaging consists in acquiring several images of a given patient with the same imaging modality (e.g. MRI, CT, US, SPECT, etc.) but with varying acquisition parameters. For instance, using Magnetic Resonance Imaging (MRI), patients followed for multiple sclerosis may undergo every six months a 3-D multisequence MR acquisition protocol with different pulse sequences (called T1, T2, PD, Flair etc): by varying some parameters of the pulse sequences (e.g Echo Time and Repetition Time), images of the same regions are produced with quite different contrasts depending on the nature and function of the observed structures. In addition, one of the acquisitions (T1) can be combined with the injection of a contrast product (typically Gadolinium) to reveal vessels and some pathologies. Diffusion tensor images (DTI) can be acquired to measure the self diffusion of protons in every voxel, allowing the measurement for instance of the direction of white matter fibers in the brain (the same principle can be used to measure the direction of muscular fibers in the heart). Functional MR images of the brain can be acquired by exploiting the so-called Bold Effect (Blood Oxygen Level Dependency): slightly higher blood flow in active regions creates a subtle higher T2* signal which can be detected with sophisticated image processing techniques.

[0]Multimodal acquisition consists in acquiring from the same patient images of different modalities, in order to exploit their complementary nature. For instance CT and MR may provide information on the anatomy (CT providing contrast between bones and soft tissues, MR providing contrast within soft tissues of different nature) while SPECT and PET images may provide functional information by measuring a local level of metabolic activity.

Despite remarkable efforts and advances during the past twenty years, the central problems of segmentation and registration have not been solved in the general case. It is our objective in the short term to work on specific versions of these problems, taking into account as much *a priori* information as possible on the underlying anatomy and pathology at hand. It is also our objective to include more knowledge of the physics of image acquisition and observed tissues, as well as of the biological processes involved. Therefore the research activities mentioned in this section will incorporate the advances made in Computational Anatomy and Computational Physiology as described in sections 3.3 and 3.4 .

We plan to pursue our efforts on the following problems:

1. Multi-dimensional, multi-sequence and multi-modal image segmentation,
2. Image Registration/Fusion,

3.3. Computational Anatomy

The objective of the Computational Anatomy (CA) is the modeling and analysis of biological variability of human anatomy. Typical applications cover the simulation of average anatomies and normal variations, the discovery of structural differences between healthy and diseased populations, and the detection and classification of pathologies from structural anomalies [0].

Studying the variability of biological shapes is an old problem (cf. the remarkable book "On Shape and Growth" by D'Arcy Thompson [69]). Significant efforts have been made since that time to develop a theory for statistical shape analysis (one can refer to [55] for a good synthesis, and to the special issue of Neuroimage [68] for recent developments). Despite all these efforts, there are a number of challenging mathematical issues which remain largely unsolved in general. A particular issue is the computation of statistics on manifolds which can be of infinite dimension (e.g the group of diffeomorphisms).

There is a classical stratification of the problems into the following 3 levels [64]: 1) construction from medical images of anatomical manifolds of points, curves, surfaces and volumes; 2) assignment of a point to point correspondence between these manifolds using a specified class of transformations (e.g. rigid, affine, diffeomorphism); 3) generation of probability laws of anatomical variation from these correspondences.

We plan to focus our efforts to the following problems:

1. Statistics on anatomical manifolds,
2. Propagation of variability from anatomical manifolds,
3. Linking anatomical variability to image analysis algorithms,
4. Grid-Computing Strategies to exploit large databases.

3.4. Computational Physiology

The objective of Computational Physiology (CP) is to provide models of the major functions of the human body and numerical methods to simulate them. The main applications are in medicine where CP can be used for instance to better understand the basic processes leading to the appearance of a pathology, to model its probable evolution and to plan, simulate, and monitor its therapy.

Quite advanced models have already been proposed to study at the molecular, cellular and organic level a number of physiological systems (see for instance [65], [62], [53], [66], [59]). While these models and new ones need to be developed, refined or validated, a grand challenge that we want to address in this project is the automatic adaptation of the model to a given patient by comparing the model with the available biomedical images and signals and possibly also some additional information (e.g. genetic). Building such *patient-specific models* is an ambitious goal which requires the choice or construction of models with a complexity adapted to the resolution of the accessible measurements and the development of new data assimilation methods coping with massive numbers of measurements and unknowns.

[0] The NIH has lauched the Alzheimer's Disease Neuroimaging Initiative (60 million USD), a multi-center MRI study of 800 patients who will be followed during several years. The objective will be to establish new surrogate end-points from the automated analysis of temporal sequences. This is a challenging objective for researchers in Computational Anatomy. The data will be made available to qualified research groups involved or not in the study.

There is a hierarchy of modeling levels for CP models of the human body [54]:

- the first level is mainly geometrical, and addresses the construction of a digital description of the anatomy [49], essentially acquired from medical imagery;
- the second level is physical, involving mainly the biomechanical modeling of various tissues, organs, vessels, muscles or bone structures [60];
- the third level is physiological, involving a modeling of the functions of the major organic systems [61] (e.g. cardiovascular, respiratory, digestive, central or peripheral nervous, muscular, reproductive, hormonal, etc.) or some pathological metabolism (e.g. evolution of cancerous or inflammatory lesions, formation of vessel stenoses, etc.);
- a fourth level would be cognitive, modeling the higher functions of the human brain [40].

These different levels of modeling are closely related to each other, and several physiological systems may interact with each other (e.g. the cardiopulmonary interaction [63]). The choice of the resolution at which each level is described is important, and may vary from microscopic to macroscopic, ideally through multiscale descriptions.

Building this complete hierarchy of models is necessary to evolve from a *Visible Human* project (essentially first level of modeling) to a much more ambitious *Physiological Human project* (see [61], [62]). We will not address all the issues raised by this ambitious project, but instead focus on topics detailed below. Among them, our objective is to identify some common methods for the resolution of the large inverse problems raised by the coupling of physiological models to medical images for the construction of patient-specific models (e.g. specific variational or sequential methods (EKF), dedicated particle filters, etc.). We also plan to develop specific expertise on the extraction of geometrical meshes from medical images for their further use in simulation procedures. Finally, computational models can be used for specific image analysis problems studied in section 3.2 (e.g. segmentation, registration, tracking, etc.). Application domains include

1. Surgery Simulation,
2. Cardiac Imaging,
3. Brain tumors, neo-angiogenesis, wound healing processes, ovocyte regulation, ...

3.5. Clinical Validation

If the objective of many of the research activities of the project is the discovery of original methods and algorithms with a demonstration of feasibility on a limited number of representative examples (i.e. proofs of concept) and publications in high quality scientific journals, we believe that it is important that a reasonable number of studies include a much more significant validation effort. As the BioMedical Image Analysis discipline becomes more mature, validation is necessary for the transformation of new ideas into clinical tools and/or industrial products. It is also often the occasion to get access to larger databases of images and signals which in turn help stimulate of new ideas and concepts.

ATHENA Project-Team

3. Research Program

3.1. Computational Diffusion MRI

Diffusion MRI (dMRI) provides a non-invasive way of estimating in-vivo CNS fiber structures using the average random thermal movement (diffusion) of water molecules as a probe. It's a recent field of research with a history of roughly three decades. It was introduced in the mid 80's by Le Bihan et al [64], Merboldt et al [68] and Taylor et al [80]. As of today, it is the unique non-invasive technique capable of describing the neural connectivity in vivo by quantifying the anisotropic diffusion of water molecules in biological tissues. The great success of dMRI comes from its ability to accurately describe the geometry of the underlying microstructure and probe the structure of the biological tissue at scales much smaller than the imaging resolution.

The diffusion of water molecules is Brownian in an isotropic medium and under normal unhindered conditions, but in fibrous structure such as white matter, the diffusion is very often directionally biased or anisotropic and water molecules tend to diffuse along fibers. For example, a molecule inside the axon of a neuron has a low probability to cross a myelin membrane. Therefore the molecule will move principally along the axis of the neural fiber. Conversely if we know that molecules locally diffuse principally in one direction, we can make the assumption that this corresponds to a set of fibers.

3.1.1. Diffusion Tensor Imaging

Shortly after the first acquisitions of diffusion-weighted images (DWI) were made in vivo [70], [71], Basser et al [45], [44] proposed the rigorous formalism of the second order Diffusion Tensor Imaging model (DTI). DTI describes the three-dimensional (3D) nature of anisotropy in tissues by assuming that the average diffusion of water molecules follows a Gaussian distribution. It encapsulates the diffusion properties of water molecules in biological tissues (inside a typical 1-3 mm^3 sized voxel) as an effective self-diffusion tensor given by a 3×3 symmetric positive definite tensor $\mathbf{D}$ [45], [44]. Diffusion tensor imaging (DTI) thus produces a three-dimensional image containing, at each voxel, the estimated tensor $\mathbf{D}$. This requires the acquisition of at least six Diffusion Weighted Images (DWI) S_k in several non-coplanar encoding directions as well as an unweighted image S_0. Because of the signal attenuation, the image noise will affect the measurements and it is therefore important to take into account the nature and the strength of this noise in all the pre-processing steps. From the diffusion tensor $\mathbf{D}$, a neural fiber direction can be inferred from the tensor's main eigenvector while various diffusion anisotropy measures, such as the Fractional Anisotropy (FA), can be computed using the associated eigenvalues to quantify anisotropy, thus describing the inequality of diffusion values among particular directions.

DTI has now proved to be extremely useful to study the normal and pathological human brain [65], [55]. It has led to many applications in clinical diagnosis of neurological diseases and disorder, neurosciences applications in assessing connectivity of different brain regions, and more recently, therapeutic applications, primarily in neurosurgical planning. An important and very successful application of diffusion MRI has been brain ischemia, following the discovery that water diffusion drops immediately after the onset of an ischemic event, when brain cells undergo swelling through cytotoxic edema.

The increasing clinical importance of diffusion imaging has drived our interest to develop new processing tools for Diffusion MRI. Because of the complexity of the data, this imaging modality raises a large amount of mathematical and computational challenges. We have therefore started to develop original and efficient algorithms relying on Riemannian geometry, differential geometry, partial differential equations and front propagation techniques to correctly and efficiently estimate, regularize, segment and process Diffusion Tensor MRI (DT-MRI) (see [67], [8] and [66]).

3.1.2. High Angular Resolution Diffusion Imaging

In DTI, the Gaussian assumption over-simplifies the diffusion of water molecules. While it is adequate for voxels in which there is only a single fiber orientation (or none), it breaks for voxels in which there are more complex internal structures. This is an important limitation, since resolution of DTI acquisition is between 1mm^3 and 3mm^3 while the physical diameter of fibers can be between 1μm and 30 μm [76], [46]. Research groups currently agree that there is complex fiber architecture in most fiber regions of the brain [75]. In fact, it is currently thought that between one third to two thirds of imaging voxels in the human brain white matter contain multiple fiber bundle crossings [47]. This has led to the development of various High Angular Resolution Diffusion Imaging (HARDI) techniques [82] such as Q-Ball Imaging (QBI) or Diffusion Spectrum Imaging (DSI) [83], [84], [86] to explore more precisely the microstructure of biological tissues.

HARDI samples q-space along as many directions as possible in order to reconstruct estimates of the true diffusion probability density function (PDF) – also referred as the Ensemble Average Propagator (EAP) – of water molecules. This true diffusion PDF is model-free and can recover the diffusion of water molecules in any underlying fiber population. HARDI depends on the number of measurements N and the gradient strength (b-value), which will directly affect acquisition time and signal to noise ratio in the signal.

Typically, there are two strategies used in HARDI: 1) sampling of the whole q-space 3D Cartesian grid and estimation of the EAP by inverse Fourier transformation or 2) single shell spherical sampling and estimation of fiber distributions from the diffusion/fiber ODF (QBI), Persistent Angular Structure [63] or Diffusion Orientation Transform [88]. In the first case, a large number of q-space points are taken over the discrete grid ($N > 200$) and the inverse Fourier transform of the measured Diffusion Weighted Imaging (DWI) signal is taken to obtain an estimate of the diffusion PDF. This is Diffusion Spectrum Imaging (DSI) [86], [83], [84]. The method requires very strong imaging gradients ($500 \leq b \leq 20000$ s/mm^2) and a long time for acquisition (15-60 minutes) depending on the number of sampling directions. To infer fiber directions of the diffusion PDF at every voxel, people take an isosurface of the diffusion PDF for a certain radius. Alternatively, they can use the second strategy known as Q-Ball imaging (QBI) i.e just a single shell HARDI acquisition to compute the diffusion orientation distribution function (ODF). With QBI, model-free mathematical approaches can be developed to reconstruct the angular profile of the diffusion displacement probability density function (PDF) of water molecules such as the ODF function which is fundamental in tractography due to the fact that it contains the full angular information of the diffusion PDF and has its maxima aligned with the underlying fiber directions at every voxel.

QBI and the diffusion ODF play a central role in our work related to the development of a robust and linear spherical harmonic estimation of the HARDI signal and to our development of a regularized, fast and robust analytical QBI solution that outperforms the state-of-the-art ODF numerical technique available. Those contributions are fundamental and have already started to impact on the Diffusion MRI, HARDI and Q-Ball Imaging community [54]. They are at the core of our probabilistic and deterministic tractography algorithms devised to best exploit the full distribution of the fiber ODF (see [52], [4] and [53],[5]).

3.1.3. High Order Tensors

Other High Order Tensors (HOT) models to estimate the diffusion function while overcoming the shortcomings of the 2nd order tensor model have also been recently proposed such as the Generalized Diffusion Tensor Imaging (G-DTI) model developed by Ozarslan et al [87], [89] or 4th order Tensor Model [43]. For more details, we refer the reader to our articles in [56], [79] where we review HOT models and to our articles in [7], co-authored with some of our close collaborators, where we review recent mathematical models and computational methods for the processing of Diffusion Magnetic Resonance Images, including state-of-the-art reconstruction of diffusion models, cerebral white matter connectivity analysis, and segmentation techniques. Recently, we started to work on Diffusion Kurtosis Imaging (DKI), of great interest for the company OLEA MEDICAL. Indeed, DKI is fast gaining popularity in the domain for characterizing the diffusion propagator or EAP by its deviation from Gaussianity. Hence it is an important tool in the clinic for characterizing the white-matter's integrity with biomarkers derived from the 3D 4th order kurtosis tensor (KT) [59].

All these powerful techniques are of utmost importance to acquire a better understanding of the CNS mechanisms and have helped to efficiently tackle and solve a number of important and challenging problems. They have also opened up a landscape of extremely exciting research fields for medicine and neuroscience. Hence, due to the complexity of the CNS data and as the magnetic field strength of scanners increase, as the strength and speed of gradients increase and as new acquisition techniques appear [3], [2], these imaging modalities raise a large amount of mathematical and computational challenges at the core of the research we develop at ATHENA [58], [79].

3.1.4. Improving dMRI Acquisitions and Modeling

One of the most important challenges in diffusion imaging is to improve acquisition schemes and analyse approaches to optimally acquire and accurately represent diffusion profiles in a clinically feasible scanning time. Indeed, a very important and open problem in Diffusion MRI is related to the fact that HARDI scans generally require many times more diffusion gradient than traditional diffusion MRI scan times. This comes at the price of longer scans, which can be problematic for children and people with certain diseases. Patients are usually unable to tolerate long scans and excessive motion of the patient during the acquisition process can force a scan to be aborted or produce useless diffusion MRI images.

Recently, we have developed novel methods for the acquisition and the processing of diffusion magnetic resonance images, to efficiently provide, with just few measurements, new insights into the structure and anatomy of the brain white matter in vivo.

First, we contributed developing real-time reconstruction algorithm based on the Kalman filter [3]. Then, and more recently, we started to explore the utility of Compressive Sensing methods to enable faster acquisition of dMRI data by reducing the number of measurements, while maintaining a high quality for the results. Compressed Sensing (CS) is a recent technique which has been proved to accurately reconstruct sparse signals from undersampled measurements acquired below the Shannon-Nyquist rate [69].

We have contributed to the reconstruction of the diffusion signal and its important features as the orientation distribution function and the ensemble average propagator, with a special focus on clinical setting in particular for single and multiple Q-shell experiments [69], [49], [50]. Compressive sensing as well as the parametric reconstruction of the diffusion signal in a continuous basis of functions such as the Spherical Polar Fourier basis, have been proved through our recent contributions to be very useful for deriving simple and analytical closed formulae for many important dMRI features, which can be estimated via a reduced number of measurements [69], [49], [50].

We have also contributed to design optimal acquisition schemes for single and multiple q-shell experiments. In particular, the method proposed in [2] helps generate sampling schemes with optimal angular coverage for multi-shell acquisitions. The cost function we proposed is an extension of the electrostatic repulsion to multi-shell and can be used to create acquisition schemes with incremental angular distribution, compatible with prematurely stopped scans. Compared to more commonly used radial sampling, our method improves the angular resolution, as well as fiber crossing discrimination. The optimal sampling schemes, freely available for download [0], have been selected for use in the HCP (Human Connectome Project) [0].

We think that such kind of contributions open new perspectives for dMRI applications including, for example, tractography where the improved characterization of the fiber orientations is likely to greatly and quickly help tracking through regions with and/or without crossing fibers [57]

3.2. MEG and EEG

Electroencephalography (EEG) and Magnetoencephalography (MEG) are two non-invasive techniques for measuring (part of) the electrical activity of the brain. While EEG is an old technique (Hans Berger, a German neuropsychiatrist, measured the first human EEG in 1929), MEG is a rather new one: the first measurements of the magnetic field generated by the electrophysiological activity of the brain were made in 1968 at MIT by

[0] http://www.emmanuelcaruyer.com/
[0] http://humanconnectome.org/documentation/Q1/imaging-protocols.html

D. Cohen. Nowadays, EEG is relatively inexpensive and is routinely used to detect and qualify neural activities (epilepsy detection and characterisation, neural disorder qualification, BCI, ...). MEG is, comparatively, much more expensive as SQUIDS only operate under very challenging conditions (at liquid helium temperature) and as a specially shielded room must be used to separate the signal of interest from the ambient noise. However, as it reveals a complementary vision to that of EEG and as it is less sensitive to the head structure, it also bears great hopes and an increasing number of MEG machines are being installed throughout the world. Inria and ODYSSÉE/ATHENA have participated in the acquisition of one such machine installed in the hospital "La Timone" in Marseille.

MEG and EEG can be measured simultaneously (M/EEG) and reveal complementary properties of the electrical fields. The two techniques have temporal resolutions of about the millisecond, which is the typical granularity of the measurable electrical phenomena that arise within the brain. This high temporal resolution makes MEG and EEG attractive for the functional study of the brain. The spatial resolution, on the contrary, is somewhat poor as only a few hundred data points can be acquired simultaneously (about 300-400 for MEG and up to 256 for EEG). MEG and EEG are somewhat complementary with fMRI and SPECT in that those provide a very good spatial resolution but a rather poor temporal resolution (of the order of a second for fMRI and a minute for SPECT). Also, contrarily to fMRI, which “only” measures an haemodynamic response linked to the metabolic demand, MEG and EEG measure a direct consequence of the electrical activity of the brain: it is acknowledged that the signals measured by MEG and EEG correspond to the variations of the post-synaptic potentials of the pyramidal cells in the cortex. Pyramidal neurons compose approximately 80% of the neurons of the cortex, and it requires at least about 50,000 active such neurons to generate some measurable signal.

While the few hundred temporal curves obtained using M/EEG have a clear clinical interest, they only provide partial information on the localisation of the sources of the activity (as the measurements are made on or outside of the head). Thus the practical use of M/EEG data raises various problems that are at the core of the ATHENA research in this topic:

- First, as acquisition is continuous and is run at a rate up to 1kHz, the amount of data generated by each experiment is huge. Data selection and reduction (finding relevant time blocks or frequency bands) and pre-processing (removing artifacts, enhancing the signal to noise ratio, ...) are largely done manually at present. Making a better and more systematic use of the measurements is an important step to optimally exploit the M/EEG data [1].
- With a proper model of the head and of the sources of brain electromagnetic activity, it is possible to simulate the electrical propagation and reconstruct sources that can explain the measured signal. Proposing better models [6], [9] and means to calibrate them [85] so as to have better reconstructions are other important aims of our work.
- Finally, we wish to exploit the temporal resolution of M/EEG and to apply the various methods we have developed to better understand some aspects of the brain functioning, and/or to extract more subtle information out of the measurements. This is of interest not only as a cognitive goal, but it also serves the purpose of validating our algorithms and can lead to the use of such methods in the field of Brain Computer Interfaces. To be able to conduct such kind of experiments, an EEG lab has been set up at ATHENA.

BAMBOO Project-Team

3. Research Program

3.1. Symbiosis

The study we propose to do on symbiosis decomposes into four main parts - (1) genetic dialog, (2) metabolic dialog, (3) symbiotic dialog and genome evolution, and (4) symbiotic dynamics - that are however strongly interrelated, and the study of such interrelations will represent an important part of our work. Another biological objective, larger and which we hope within the ERC project SISYPHE just to sketch for a longer term investigation, will aim at getting at a better grasp of species identity and of a number of identity-related concepts. We now briefly indicate the main points that have started been investigated or should be investigated in the next five years.

Genetic dialog
We plan to study the genetic dialog at the regulation level between symbiont and host by addressing the following mathematical and algorithmic issues:

1. model and identify all small RNAs from the bacterium and the host which may be involved in the genetic dialog between the two, and model/identify the targets of such small RNAs;
2. infer selected parts of the regulatory network of both symbiont and host (this will enable to treat the next point) using all available information;
3. explore at both the computational and experimental levels the complementarity of the two networks, and revisit at a network level the question of a regulatory response of the symbiont to its host's demand;
4. compare the complementarities observed between pairs of networks (the host's and the symbiont's); such complementarities will presumably vary with the different types of host-symbiont relationships considered, and of course with the information the networks model (structural or dynamic); Along the way, it may become important at some point to address also the issue of transposable elements (abbreviated into TEs, that are genes which can jump spontaneously from one site to another in a genome following or not a duplication event). It is increasingly believed that TEs play a role in the regulation of the expression of the genes in eukaryotic genomes. The same role in symbionts, and in the host-symbiont dialog has been less or not explored. This requires to address the following additional task:
5. accurately and systematically detect all transposable elements (*i.e.* genes which can jump spontaneously from one site to another in a genome following or not a duplication event) and assess their implication in their own regulation and that of their host genome (the new sequencing technologies should facilitate this task as well as other data expression analyses, if we are able to master the computational problem of analysing the flow of data they generate: fragment indexing, mapping and assembly);
6. where possible, obtain data enabling to infer the PPI (Protein-Protein Interaction) for hosts and symbionts, and at the host-symbiont interface and analyse the PPI networks obtained and how they interact.

Initial algorithmic and statistical approaches for the first two items above are under way and are sustained by a well-established expertise of the team on sequence and microarray bioinformatic analysis. Both problems are however notoriously hard because of the high level of missing data and noise, and of our relative lack of knowledge of what could be the key elements of genetic regulation, such as small and micro RNAs.

We also plan to establish the complete repertoire of transcription factors of the interacting partners (with possible exchanges between them) at both the computational and experimental levels. Comparative biology (search by sequence homology of known regulators), 3D-structural modelling of putative domains interacting with the DNA molecule, regulatory domains conserved in the upstream region of coding DNA are among classical and routinely used methods to search for putative regulatory proteins and elements in the genomes. Experimentally, the BiaCore (using the surface plasmon resonance principle) and ChIP-Seq (using chromatin precipitation coupled with high-throughput sequencing from Solexa) techniques offer powerful tools to capture all the protein-DNA interactions corresponding to a specific putative regulator. However, these techniques have not been evaluated in the context of interacting partners making this task an interesting challenge.

Metabolic dialog

Our main plan for this part, where we have already many results, some obtained this last year, is to:

1. continue with and improve our work on reconstructing the metabolic networks of organisms with sequenced genomes, taking in particular care to cover as much as possible the different types of hosts and symbionts in interaction;
2. refine the network reconstructions by using flux balance analysis which will in turn require addressing the next item;
3. improve our capacity to efficiently compute fluxes and do flux balance analysis; current algorithms can handle only relatively small networks;
4. analyse and compare the networks in terms of their general structural, quantitative and dynamic characteristics;
5. develop models and algorithms to compare different types of metabolic interfaces which will imply being able, by a joint computational and experimental approach, to determine what is transported across interacting metabolisms;
6. define what would be a good null hypothesis to test the statistical significance, and therefore possible biological relevance of the characteristics observed when analysing or comparing (random network problem, a mostly open issue despite the various models available);
7. use the results from item 5, that is indications on the precursors of a bacterial metabolism that are key players in the dialog with the metabolism of the host, to revisit the genetic regulation dialog between symbiont and host.

Computational results from the last item will be complemented with experiments to help understand what is transported from the host to the symbiont and how what is transported may be related with the genetic dialog between the two organisms (items 5 and 6).

Great care will also be taken in all cases (metabolism- or regulation-only, or both together) to consider the situations, rather common, where more than two partners are involved in a symbiosis, that is when there are secondary symbionts of a same host.

The first five items above have started being computationally explored by our team, as has the last item including experimentally. Some algorithmic proofs-of-concept, notably as concerns structural, flux, precursor and chemical organisation studies (see some of the publications of the last year and this one), have been established but much more work is necessary. The main difficulties with items 3 and 4 are of two sorts. The first one is a modelling issue: what are the best models for analysing and comparing two or more networks? This will greatly depend on the biological question put, whether evolutionary or functional, structural or physiologic, besides being a choice that should be motivated by the extent and quality of the data available. The second sort of difficulty ,which also applies to other items notably (item 2), is computational. Most of the problems related with analysing and specially comparing are known to be hard but many issues remain open. The question of a good random model (item 6) is also largely open.

Symbiotic dialog and genome evolution

Genomes are not static. Genes may get duplicated, sometimes the duplication affects the whole genome, or genes can transpose, while whole genomic segments can be reversed or deleted. Deletions are indeed one of the most common events observed for some symbionts. Genetic material may also be transferred across sub-species or species (lateral transfer), thus leading to the insertion of new elements in a genome. Finally, parts of a genome may be amplified through, for instance, slippage during DNA replication resulting in the multiplication of the copies of a repeat that appear tandemly arrayed along a genome. Tandem repeats, and other types of short or long repetitions are also believed to play a role in the generation of new genomic rearrangements although whether they are always the cause or consequence of the genome break and gene order change remains a disputed issue.

Work on this part will involve the following items:

1. extend the theoretical work done in the past years (rearrangement distance, rearrangement scenarios enumeration) to deal with different types of rearrangements and explore various types of biological constraints;
2. develop good random models (a largely open question despite some initial work in the area) for rearrangement distances and scenarios under a certain model, i.e. type of rearrangement operation(s) and of constraint(s), to assess whether the distances / scenarios observed have statistically notable characteristics;
3. extensively use the method(s) developed to investigate the rearrangement histories for the families of symbionts whose genomes have been sequenced and sufficiently annotated;
4. investigate the correlation of such histories with the repeats content and distribution along the genomes;
5. use the results of the above analyses together with a natural selection criterion to revisit the optimality model of rearrangement dynamics;
6. extend such model to deal with eukaryotic (multi-chromosomal) genomes;
7. at the interface host-symbiont, investigate the relation between the rearrangement histories in hosts and symbionts and the various types of symbiotic relationships observed in nature;
8. map such histories and their relation with the genetic and metabolic networks of hosts and symbionts, separately and at the interface;
9. develop methods to identify and quantify rearrangement events from NGS data.

Symbiotic dynamics

In order to understand the evolutionary consequences of symbiotic relations and their long term trajectories, one should be able to assess how tight is the association between symbionts and their hosts.

The main questions we would like to address are:

1. how often are symbionts horizontally transferred among branches of the host phylogenetic tree?
2. how long do parasites persist inside their host following the invasion of a new lineage?
3. what processes underlie this dynamic gain/loss equilibrium?

Mathematically, these questions have been traditionally addressed by co-phylogenetic methods, that is by comparing the evolutionary histories of hosts and parasites as represented in phylogenetic trees.

Currently available co-phylogenetic algorithms present various types of limitations as suggested in recent surveys. This may seriously compromise their interpretation with a view to understanding the evolutionary dynamics of parasites in communities. A few examples of limitations are the (often wrong) assumption made that the same rates of loss and gain of parasite infection apply for every host taxonomic group, and the fact that the possibility of multi-infections is not considered. In the latter case, exchange of genetic material between different parasites of a same host could further scramble the co-evolutionary signal. We therefore plan to:

1. better formalise the problem and the different simplifications that could be made, or inversely, should be avoided in the co-phylogeny studies; examples of the latter are the possibility of multi-infections, differential rate of loss and gain of infection depending on the host taxonomic group and geographic distance between hosts, etc., and propose better co-phylogenetic algorithms;
2. elaborate series of simulated data that will enable to (i) get a better grasp of the effect of the different parameters of the problem and, more practically, (ii) evaluate the performance of the method(s) that exist or are proposed (see next item);
3. apply the new methods to address the three questions above.

3.2. Intracellular interactions

The interactions of a symbiont with others sharing a same host, or with a symbiont and the cell of its host in the case of endosymbionts (organism that lives within the body or cells of another) are special, perhaps more complex cases of intracellular interactions that may concern different types of genetic elements, from organelles to whole chromosomes. The spatial arrangement of those genetic elements inside the nucleus of a cell is believed to be important both for gene expression and exchanges of genetic material between chromosomes. This question goes beyond the symbiosis one and has been investigated in the team in the last few years. Work on this will continue in future and concern developing algorithmic and statistical methods to analyse the interaction data that is starting to become available, in particular using NGS methods, in order to arrive at a better understanding of transcription, regulation both classical and epigenetic (inherited changes in phenotype or gene expression caused by mechanisms other than changes in the underlying DNA sequence), alternative splicing and trans-splicing phenomena, as well as study the possible interactions between an eukaryotic cell and its organelles or other cytoplasmic structures.

BEAGLE Project-Team

3. Research Program

3.1. Introduction

As stated above, the research topics of the BEAGLE Team are centered on the modelisation and simulation of cellular processes. More specifically, we focus on two specific processes that govern cell dynamics and behavior: Evolution and Biophysics. This leads to two main topics: computational cell biology and models for genome evolution.

3.2. Computational Cell Biology

BEAGLE contributes computational models and simulations to the study of cell signaling in prokaryotic and eukaryotic cells, with a special focus on the dynamics of cell signaling both in time and in space. Importantly, our objective here is not so much to produce innovative computer methodologies, but rather to improve our knowledge of the field of cell biology by means of computer methodologies.

This objective is not accessible without a thorough immersion in experimental cell biology. Hence, one specificity of BEAGLE is to be closely associated inside each research project with experimental biology groups. For instance, all the current PhD students implicated in the research projects below have strong interactions with experimenters, most of them conducting experiments themselves in our collaborators' labs. In such a case, the supervision of their PhD is systematically shared between an experimentalist and a theoretician (modeler/computer scientist).

Standard modeling works in cell biochemistry are usually based on mean-field equations, most often referred to as "laws of mass-action". Yet, the derivation of these laws is based on strict assumptions. In particular, the reaction medium must be dilute, perfectly-mixed, three-dimensional and spatially homogeneous and the resulting kinetics are purely deterministic. Many of these assumptions are obviously violated in cells. As already stressed out before, the external membrane or the interior of eukaryotic as well as prokaryotic cells evidence spatial organization at several length scales, so that they must be considered as non-homogeneous media. Moreover, in many case, the small number of molecule copies present in the cell violates the condition for perfect mixing, and more generally, the "law of large numbers" supporting mean-field equations.

When the laws-of-mass-action are invalidated, individual-based models (IBM) appear as the best modeling alternative to evaluate the impact of these specific cellular conditions on the spatial and temporal dynamics of the signaling networks. We develop Individual-Based Models to evaluate the fundamental impact of non-homogeneous space conditions on biochemical diffusion and reaction. More specifically, we focus on the effects of two major sources of non-homogeneity within cells: macromolecular crowding and non-homogeneous diffusion. Macromolecular crowding provides obstacles to the diffusive movement of the signaling molecules, which may in turn have a strong impact on biochemical reactions [35]. In this perspective, we use IBM to renew the interpretation of the experimental literature on this aspect, in particular in the light of the available evidence for anomalous subdiffusion in living cells. Another pertinent source of non-homogeneity is the presence of lipid rafts and/or caveolae in eukaryotic cell membranes that locally alter diffusion. We showed several properties of these diffusion gradients on cells membranes. In addition, combining IBMs and cell biology experiments, we investigate the spatial organization of membrane receptors in plasmic membranes and the impact of these spatial features on the initiation of the signaling networks [39]. More recently, we started to develop IBMs to propose experimentally-verifiable tests able to distinguish between hindered diffusion due to obstacles (macromolecular crowding) and non-homogeneous diffusion (lipid rafts) in experimental data.

The last aspect we tackle concerns the stochasticity of gene expression. Indeed, the stochastic nature of gene expression at the single cell level is now a well established fact [45]. Most modeling works try to explain this stochasticity through the small number of copies of the implicated molecules (transcription factors, in particular). In collaboration with the experimental cell biology group led by Olivier Gandrillon at the Centre de Génétique et de Physiologie Moléculaire et Cellulaire (CGPhyMC, UMR CNRS 5534), Lyon, we study how stochastic gene expression in eukaryotic cells is linked to the physical properties of the cellular medium (e.g., nature of diffusion in the nucleoplasm, promoter accessibility to various molecules, crowding). We have already developed a computer model whose analysis suggests that factors such as chromatin remodeling dynamics have to be accounted for [41]. Other works introduce spatial dimensions in the model, in particular to estimate the role of space in complex (protein+ DNA) formation. Such models should yield useful insights into the sources of stochasticity that are currently not explained by obvious causes (e.g. small copy numbers).

3.3. Models of genome evolution

Classical artificial evolution frameworks lack the basic structure of biological genome (i.e. a double-strand sequence supporting variable size genes separated by variable size intergenic sequences). Yet, if one wants to study how a mutation-selection process is likely (or not) to result in particular biological structures, it is mandatory that the effect of mutation modifies this structure in a realistic way. We have developed an artificial chemistry based on a mathematical formulation of proteins and of the phenotypic traits. In our framework, the digital genome has a structure similar to prokaryotic genomes and a non-trivial genotype-phenotype map. It is a double-stranded genome on which genes are identified using promoter-terminator- like and start-stop-like signal sequences. Each gene is transcribed and translated into an elementary mathematical element (a "protein") and these elements – whatever their number – are combined to compute the phenotype of the organism. The Aevol (Artificial EVOLution) model is based on this framework and is thus able to represent genomes with variable length, gene number and order, and with a variable amount of non-coding sequences (for a complete description of the model, see [52]).

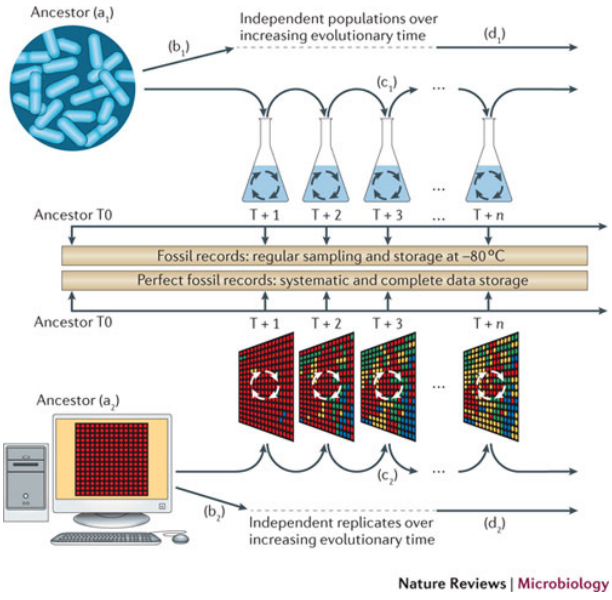


Figure 1. Parallel between experimental evolution and artificial evolution

As a consequence, this model can be used to study how evolutionary pressures like the ones for robustness or evolvability can shape genome structure [53], [50], [51], [60]. Indeed, using this model, we have shown that genome compactness is strongly influenced by indirect selective pressures for robustness and evolvability. By genome compactness, we mean several structural features of genome structure, like gene number, amount of non functional DNA, presence or absence of overlapping genes, presence or absence of operons [53], [50], [61]. More precisely, we have shown that the genome evolves towards a compact structure if the rate of spontaneous mutations and rearrangements is high. As far as gene number is concerned, this effect was known as an error-threshold effect [44]. However, the effect we observed on the amount of non functional DNA was unexpected. We have shown that it can only be understood if rearrangements are taken into account: by promoting large duplications or deletions, non functional DNA can be mutagenic for the genes it surrounds.

We have extended this framework to include genetic regulation (R-Aevol variant of the model). We are now able to study how these pressures also shape the structure and size of the genetic network in our virtual organisms [37], [36], [38]. Using R-Aevol we have been able to show that (i) the model qualitatively reproduces known scaling properties in the gene content of prokaryotic genomes and that (ii) these laws are not due to differences in lifestyles but to differences in the spontaneous rates of mutations and rearrangements [36]. Our approach consists in addressing unsolved questions on Darwinian evolution by designing controlled and repeated evolutionary experiments, either to test the various evolutionary scenarios found in the literature or to propose new ones. Our experience is that “thought experiments” are often misleading: because evolution is a complex process involving long-term and indirect effects (like the indirect selection of robustness and evolvability), it is hard to correctly predict the effect of a factor by mere thinking. The type of models we develop are particularly well suited to provide control experiments or test of null hypotheses for specific evolutionary scenarios. We often find that the scenarios commonly found in the literature may not be necessary, after all, to explain the evolutionary origin of a specific biological feature. No selective cost to genome size was needed to explain the evolution of genome compactness [53], and no difference in lifestyles and environment was needed to explain the complexity of the gene regulatory network [36]. When we unravel such phenomena in the individual-based simulations, we try to build ”simpler” mathematical models (using for instance population genetics-like frameworks) to determine the minimal set of ingredients required to produce the effect. Both approaches are complementary: the individual-based model is a more natural tool to interact with biologists, while the mathematical models contain fewer parameters and fewer ad-hoc hypotheses about the cellular chemistry.

At this time, simulating the evolution of large genomes during hundreds of thousands of generation with the Aevol software can take several weeks or even months. It is worse with Raevol, where we not only simulate mutations and selection at the evolutionary timescale, but also simulate the lifetime of the individuals, allowing them to respond to environmental signals. Previous efforts to parallelize and distribute Aevol had yielded limited results due to the lack of dedicated staff on these problems. Since September, we have started to study how to improve the performance of (R-)Aevol. Thanks to the ADT Aevol, one and a half full time engineers will work to improve Aevol and especially to parallelize it. Moreover, we are working to formalize the numerical computation problems with (R-)Aevol to use state-of-the-art optimization techniques from the HPC community. It ranges from dense and sparse matrix multiplication and their optimizations (such as Tridiagonal matrix algorithm) to using new generation accelerator (Intel Xeon Phi and NVidia Telsa). However, our goal is not to become a HPC nor a numerical computation team but to work with well-established teams in these fields, such as through the Joint Laboratory for Extreme-Scale Computing, but also with Inria teams in these fields (e.g. ROMA, Avalon, CORSE, RUNTIME, MESCAL). By doing so, (R-)Aevol simulations will be faster, allowing us to study more parameters in a shorter time. Furthermore, we will also be able to simulate more realistic population sizes, that currently do not fit into the memory of a single computer.

Little has been achieved concerning the validation of these models, and the relevance of the observed evolutionary tendencies for living organisms. Some comparisons have been made between Adiva and experimental evolution [54], [48], but the comparison with what happened in a long timescale to life on earth is still missing. It is partly because the reconstruction of ancient genomes from the similarities and differences between extant ones is a difficult computational problem which still misses good solutions for every type of mutations, in particular the ones concerning changes in the genome structure.

There exist good phylogenic models of punctual mutations on sequences [46], which enable the reconstruction of small parts of ancestral sequences, individual genes for example [55]. But models of whole genome evolution, taking into account large scale events like duplications, insertions, deletions, lateral transfer, rearrangements are just being developped [63], [42]. Integrative phylogenetic models, considering both nucleotide subsitutions and genome architectures, like Aevol does, are still missing.

Partial models lead to evolutionary hypotheses on the birth and death of genes [43], on the rearrangements due to duplications [33], [62], on the reasons of variation of genome size [49], [56]. Most of these hypotheses are difficult to test due to the difficulty of *in vivo* evolutionary experiments.

To this aim, we develop evolutionary models for reconstructing the history of organisms from the comparison of their genome, at every scale, from nucleotide substitutions to genome organisation rearrangements. These models include large-scale duplications as well as loss of DNA material, and lateral gene transfers from distant species. In particular we have developed models of evolution by rearrangements [57], methods for reconstructing the organization of ancestral genomes [58], [40], [59], or for detecting lateral gene transfer events [32], [8]. It is complementary with the Aevol development because both the model of artificial evolution and the phylogenetic models we develop emphasize on the architecture of genomes. So we are in a good position to compare artificial and biological data on this point.

We improve the phylogenetic models to reconstruct ancestral genomes, jointly seen as gene contents, orders, organizations, sequences. It will necessitate integrative models of genome evolution, which is desirable not only because they will provide a unifying view on molecular evolution, but also because they will put into light the relations between different kinds of mutations, and enable the comparison with artificial experiments from Aevol.

Based on this experience, the BEAGLE team contributes individual-based and mathematical models of genome evolution, in silico experiments as well as historical reconstruction on real genomes, to shed light on the evolutionary origin of the complex properties of cells.

BIGS Project-Team

3. Research Program

3.1. Online data analysis

Participants: J.-M. Monnez, P. Vallois. Generally speaking, there exists an overwhelming amount of articles dealing with the analysis of high dimensional data. Indeed, this is one of the major challenges in statistics today, motivated by internet or biostatistics applications. Within this global picture, the problem of classification or dimension reduction of online data can be traced back at least to a seminal paper by Mac Queen [53], in which the k-means algorithm is introduced. This popular algorithm, constructed for classification purposes, consists in a stepwise updating of the centers of some classes according to a stream of data entering into the system. The literature on the topic has been growing then rapidly since the beginning of the 90's.

Our point of view on the topic relies on the so-called *French data analysis school*, and more specifically on Factorial Analysis tools. In this context, it was then rapidly seen that stochastic approximation was an essential tool (see Lebart's paper [50]), which allows one to approximate eigenvectors in a stepwise manner. A systematic study of Principal Component and Factorial Analysis has then been lead by Monnez in the series of papers [56], [54], [55], in which many aspects of convergences of online processes are analyzed thanks to the stochastic approximation techniques.

3.2. Local regression techniques

Participants: S. Ferrigno, A. Muller-Gueudin. In the context where a response variable Y is to be related to a set of regressors X, one of the general goals of Statistics is to provide the end user with a model which turns out to be useful in predicting Y for various values of X. Except for the simplest situations, the determination of a good model involves many steps. For example, for the task of predicting the value of Y as a function of the covariate X, statisticians have elaborated models such as the regression model with random regressors:

$$Y = g(X, \theta) + \sigma(X)\epsilon.$$

Many assumptions must be made to reach it as a possible model. Some require much thinking, as for example, those related to the functional form of $g(\cdot, \theta)$. Some are made more casually, as often those related to the functional form of $\sigma(\cdot)$ or those concerning the distribution of the random error term ϵ. Finally, some assumptions are made for commodity. Thus the need for methods that can assess if a model is concordant with the data it is supposed to adjust. The methods fall under the banner of goodness of fit tests. Most existing tests are *directional*, in the sense that they can detect departures from only one or a few aspects of a null model. For example, many tests have been proposed in the literature to assess the validity of an entertained structural part $g(\cdot, \theta)$. Some authors have also proposed tests about the variance term $\sigma(\cdot)$ (cf. [51]). Procedures testing the normality of the ϵ_i are given, but for other assumptions much less work has been done. Therefore the need of a global test which can evaluate the validity of a global structure emerges quite naturally.

With these preliminaries in mind, let us observe that one quantity which embodies all the information about the joint behavior of (X, Y) is the cumulative conditional distribution function, defined by

$$F(y|x) = P(Y \leq y|X = x).$$

The (nonparametric) estimation of this function is thus of primary importance. To this aim, notice that modern estimators are usually based on the local polynomial approach, which has been recognized as superior to classical estimates based on the Nadaraya-Watson approach, and are as good as the recent versions based on spline and other methods. In some recent works [41], [42], we address the following questions:

- Construction of a global test by means of Cramér-von Mises statistic.
- Optimal bandwidth of the kernel used for approximation purposes.

We also obtain sharp estimates on the conditional distribution function in [43].

3.3. Stochastic modeling for complex and biological systems

Participants: R. Azaïs, T. Bastogne, C. Lacaux, A. Muller-Gueudin, S. Tindel, P. Vallois, S. Wantz-Mézières

In most biological contexts, mathematics turn out to be useful in producing accurate models with dual objectives: they should be simple enough and meaningful for the biologist on the one hand, and they should provide some insight on the biological phenomenon at stake on the other hand. We have focused on this kind of issue in various contexts that we shall summarize below.

Photodynamic Therapy: Photodynamic therapy induces a huge demand of interconnected mathematical systems, among which we have studied recently the following ones:

• The tumor growth model is of crucial importance in order to understand the behavior of the whole therapy. We have considered the tumor growth as a stochastic equation, for which we have handled the problem uncertainties on the measure times [27] as well as mixed effects for parameter estimation.

• Another important aspect to quantify for photodynamic therapy calibration is the response to radiotherapy treatments. There are several valid mathematical ways to describe this process, among which we distinguish the so-called hit model. This model assumes that whenever a group of sensitive targets (chromosomes, membrane) in the cell are reached by a sufficient number of radiations, then the cell is inactivated and dies. We have elaborated on this scheme in order to take into account two additional facts: (i) The reduction of the cell situation to a two-state model might be an oversimplification. (ii) Several doses of radiations are inoculated as time passes. These observations have lead us to introduce a new model based on multi-state Markov chains arguments (Keinj & al, 2012), in which cell proliferation can be incorporated.

Bacteriophage therapy: Let us mention a starting collaboration between BIGS and the Genetics and Microbiology department at the Universitat Autònoma de Barcelona, on the modeling of bacteriophage therapies. The main objective here is to describe how a certain family of benign viruses is able to weaken a bacterium induced disease, which naturally leads to the introduction of a noisy predator-prey system of equations. It should be mentioned that some similar problems have been treated (in a rather informal way, invoking a linearization procedure) by Carletti in [34]. These tools cannot be applied directly to our system, and our methods are based on concentration and large deviations techniques (on which we already had an expertise [57], [60]) in order to combine convergence to equilibrium for the deterministic system and deviations of the stochastic system. Notice that A. Muller-Gueudin is also working with A. Debussche and O. Radulescu on a related topic [37], namely the convergence of a model of cellular biochemical reactions.

Gaussian signals: Nature provides us with many examples of systems such that the observed signal has a given Hölder regularity, which does not correspond to the one we might expect from a system driven by ordinary Brownian motion. This situation is commonly handled by noisy equations driven by Gaussian processes such as fractional Brownian motion or (in higher dimensions of the parameter) fractional fields.

The basic aspects of differential equations driven by a fractional Brownian motion (fBm) and other Gaussian processes are now well understood, mainly thanks to the so-called *rough paths* tools [52], but also invoking the Russo-Vallois integration techniques [59]. The specific issue of Volterra equations driven by fBm, which is central for the subdiffusion within proteins problem, is addressed in [38].

Fractional fields are very often used to model irregular phenomena which exhibit a scale invariance property, fractional Brownian motion being the historical fractional model. Nevertheless, its isotropy property is a serious drawback for instance in hydrology or in medecine (see [33]). Moreover, the fractional Brownian motion cannot be used to model some phenomena for which the regularity varies with time. Hence, many generalizations (Gaussian or not) of this model have been recently proposed, see for instance [28] for some Gaussian locally self-similar fields, [46] for some non-Gaussian models, [31] for anisotropic models.

Our team has thus contributed [36], [47], [46], [48], [58] and still contributes [30], [32], [31], [49], [44] to this theoretical study: Hölder continuity, fractal dimensions, existence and uniqueness results for differential equations, study of the laws to quote a few examples. As we shall see below, this line of investigation also has some impact in terms of applications: we shall discuss how we plan to apply our results to osteoporosis on the one hand and to fluctuations within protein molecules on the other hand.

3.4. Parameter identifiability and estimation

Participants: R. Azaïs, T. Bastogne, S. Tindel, P. Vallois, S. Wantz-Mézières

When one desires to confront theoretical probabilistic models with real data, statistical tools are obviously crucial. We have focused on two of them: parameter identifiability and parameter estimation.

Parameter identifiability [62] deals with the possibility to give a unique value to each parameter of a mathematical model structure in inverse problems. There are many methods for testing models for identifiability: Laplace transform, similarity transform, Taylor series, local state isomorphism or elimination theory. Most of the current approaches are devoted to *a priori* identifiability and are based on algebraic techniques. We are particularly concerned with *a posteriori* identifiability, *i.e.*, after experiments or in a constrained experimental framework and the link with experimental design techniques. Our approach is based on statistical techniques through the use of variance-based methods. These techniques are strongly connected with global sensitivity approaches and Monte Carlo methods.

The parameter estimation for a family of probability laws has a very long story in statistics, and we refer to [29] for an elegant overview of the topic. Moving to the references more closely related to our specific projects, let us recall first that the mathematical description of photodynamic therapy can be split up into three parametric models : the uptake model (pharmacokinetics of the photosensitizing drug into cancer cells), the photoreaction model and the tumor growth model. Several papers have been reported for the application of system identification techniques to pharmacokinetics modeling problems. But two issues were ignored in these previous works: presence of timing noise and identification from longitudinal data. In [27], we have proposed a bounded-error estimation algorithm based on interval analysis to solve the parameter estimation problem while taking into consideration uncertainty on observation time instants. Statistical inference from longitudinal data based on mixed effects models can be performed by the *Monolix* software (http://www.lixoft.eu/products/monolix/product-monolix-overview/) developed by the Monolix group chaired by Marc Lavielle and France Mentré, and supported by Inria. In the recent past, we have used this tool for tumor growth modeling. According to what we know so far, no parameter estimation study has been reported about the photoreaction model in photodynamic therapy. A photoreaction model, composed of six stochastic differential equations, is proposed in [39]. The main open problem is to access to data. We currently build on an experimental platform which aims at overcoming this technical issue. Moreover, an identifiability study coupled to a global sensitivity analysis of the photoreaction model are currently in progress. Tumor growth is generally described by population dynamics models or by cell cycle models. Faced with this wide variety of descriptions, one of the main open problems is to identify the suitable model structure. As mentioned above, we currently investigate alternative representations based on branching processes and Markov chains, with a model selection procedure in mind.

A few words should be said about the existing literature on statistical inference for diffusion or related processes, a topic which will be at the heart of three of our projects (namely photodynamic and bacteriophage therapies, as well as fluctuations within molecules). The monograph [45] is a good reference on the basic estimation techniques for diffusion processes. The problem of estimating diffusions observed at discrete times,

of crucial importance for applications, has been addressed mainly since the mid 90s. The maximum likelihood techniques, which are also classical for parameter estimation, are well represented by the contributions [40].

Some attention has been paid recently to the estimation of the coefficients of fractional or multifractional Brownian motion according to a set of observations. Let us quote for instance the nice surveys [26], [35]. On the other hand, the inference problem for diffusions driven by a fractional Brownian motion is still in its infancy. A good reference on the question is [61], dealing with some very particular families of equations, which do not cover the cases of interest for us.

BIOCORE Project-Team

3. Research Program

3.1. Mathematical and computational methods

BIOCORE's action is centered on the mathematical modeling of biological systems, more particularly of artificial ecosystems, that have been built or strongly shaped by human. Indeed, the complexity of such systems where life plays a central role often makes them impossible to understand, control, or optimize without such a formalization. Our theoretical framework of choice for that purpose is Control Theory, whose central concept is "the system", described by state variables, with inputs (action on the system), and outputs (the available measurements on the system). In modeling the ecosystems that we consider, mainly through ordinary differential equations, the state variables are often population, substrate and/or food densities, whose evolution is influenced by the voluntary or involuntary actions of man (inputs and disturbances). The outputs will be some product that one can collect from this ecosystem (harvest, capture, production of a biochemical product, etc), or some measurements (number of individuals, concentrations, etc). Developing a model in biology is however not straightforward: the absence of rigorous laws as in physics, the presence of numerous populations and inputs in the ecosystems, most of them being irrelevant to the problem at hand, the uncertainties and noise in experiments or even in the biological interactions require the development of dedicated techniques to identify and validate the structure of models from data obtained by or with experimentalists.

Building a model is rarely an objective in itself. Once we have checked that it satisfies some biological constraints (eg. densities stay positive) and fitted its parameters to data (requiring tailor-made methods), we perform a mathematical analysis to check that its behavior is consistent with observations. Again, specific methods for this analysis need to be developed that take advantage of the structure of the model (eg. the interactions are monotone) and that take into account the strong uncertainty that is linked to life, so that qualitative, rather than quantitative, analysis is often the way to go.

In order to act on the system, which often is the purpose of our modeling approach, we then make use of two strong points of Control Theory: 1) the development of observers, that estimate the full internal state of the system from the measurements that we have, and 2) the design of a control law, that imposes to the system the behavior that we want to achieve, such as the regulation at a set point or optimization of its functioning. However, due to the peculiar structure and large uncertainties of our models, we need to develop specific methods. Since actual sensors can be quite costly or simply do not exist, a large part of the internal state often needs to be re-constructed from the measurements and one of the methods we developed consists in integrating the large uncertainties by assuming that some parameters or inputs belong to given intervals. We then developed robust observers that asymptotically estimate intervals for the state variables [91]. Using the directly measured variables and those that have been obtained through such, or other, observers, we then develop control methods that take advantage of the system structure (linked to competition or predation relationships between species in bioreactors or in the trophic networks created or modified by biological control).

3.2. A methodological approach to biology: from genes to ecosystems

One of the objectives of BIOCORE is to develop a methodology that leads to the integration of the different biological levels in our modeling approach: from the biochemical reactions to ecosystems. The regulatory pathways at the cellular level are at the basis of the behavior of the individual organism but, conversely, the external stresses perceived by the individual or population will also influence the intracellular pathways. In a modern "systems biology" view, the dynamics of the whole biosystem/ecosystem emerge from the interconnections among its components, cellular pathways/individual organisms/population. The different scales of size and time that exist at each level will also play an important role in the behavior of the biosystem/ecosystem. We intend to develop methods to understand the mechanisms at play at each level,

from cellular pathways to individual organisms and populations; we assess and model the interconnections and influence between two scale levels (eg., metabolic and genetic; individual organism and population); we explore the possible regulatory and control pathways between two levels; we aim at reducing the size of these large models, in order to isolate subsystems of the main players involved in specific dynamical behaviors.

We develop a theoretical approach of biology by simultaneously considering different levels of description and by linking them, either bottom up (scale transfer) or top down (model reduction). These approaches are used on modeling and analysis of the dynamics of populations of organisms; modeling and analysis of small artificial biological systems using methods of systems biology; control and design of artificial and synthetic biological systems, especially through the coupling of systems.

The goal of this multi-level approach is to be able to design or control the cell or individuals in order to optimize some production or behavior at higher level: for example, control the growth of microalgae via their genetic or metabolic networks, in order to optimize the production of lipids for bioenergy at the photobioreactor level.

BONSAI Project-Team

3. Research Program

3.1. Combinatorial discrete models and algorithms

Our research is driven by biological questions. At the same time, we have in mind to develop well-founded models and efficient algorithms. Biological macromolecules are naturally modelled by various types of discrete structures: String, trees, and graphs. String algorithms is an established research subject of the team. We have been working on spaced seed techniques for several years. Members of the team also have a strong expertise in text indexing and compressed index data structures, such as BWT. Such methods are widely-used for the analysis of biological sequences because they allow a data set to be stored and queried efficiently. Ordered trees and graphs naturally arise when dealing with structures of molecules, such as RNAs or non-ribosomal peptides. The underlying questions are: How to compare molecules at structural level, how to search for structural patterns ? String, trees and graphs are also useful to study genomic rearrangements: Neighborhoods of genes can be modelled by oriented graphs, genomes as permutations, strings or trees. High-performance computing is another tool that we use to achieve our goals.

3.2. Discrete statistics and probability

At a lower level, our work relies on a basic background on discrete statistics and probability. When dealing with large input data sets, it is essential to be able to discriminate between noisy features observed by chance from those that are biologically relevant. The aim here is to introduce a probabilistic model and to use sound statistical methods to assess the significance of some observations about these data. Examples of such observations are the length of a repeated region, the number of occurrences of a motif (DNA or RNA), the free energy of a conserved RNA secondary structure, etc. Probabilistic models are also used to describe genome evolution. In this context, Bayesian models and MCMC sampling allow to approximate probability distributions over free parameters and to describe biologically relevant models.

CARMEN Team

3. Research Program

3.1. Complex models for the propagation of cardiac action potentials

Cardiac arrythmias originates from the multiscale organisation of the cardiac action potential from the cellular scale up to the scale of the body. It relates the molecular processes from the cell membranes to the electrocardiogram, an electrical signal on the torso. The spatio-temporal patterns of this propagation is related both to the function of the cellular membrane and of the structural organisation of the cells into tissues, into the organ and final within the body.

Several improvements of current models of the propagation of the action potential will be developped, based on previous work [8] and on the data available at the LIRYC:

- Enrichment of the current monodomain and bidomain models by accounting for structural heterogeneities of the tissue at an intermediate scale. Here we focus on multiscale analysis techniques applied to the various high-resolution structural data available at the LIRYC.
- Coupling of the tissues from the different cardiac compartments and conduction systems. Here, we want to develop model that couples 1D, 2D and 3D phenomena described by reaction-diffusion PDEs.

These models are essential to improve our in-depth understanding of cardiac electrical dysfunction. To this aim, we will use high-performance computing techniques in order to explore numerically the complexity of these models and check that they are reliable experimental tools.

3.2. Simplified models and inverse problems

The medical and clinical exploration of the electrical signals is based on accurate reconstruction of the typical patterns of propagation of the action potential. The correct detection of these complex patterns by non-invasive electrical imaging techniques has to be developped. Both problems involve solving inverse problems that cannot be addressed with the more compex models. We want both to develop simple and fast models of the propagation of cardiac action potentials and improve the solutions to the inverse problems found in cardiac electrical imaging techniques.

The cardiac inverse problem consists in finding the cardiac activation maps or, more generally the whole cardiac electrical activity, from high density body surface electrocardiograms. It is a new and a powerful diagnosis technique, which success would be considered as a breakthrough in the cardiac diagnosis. Although widely studied during the last years, it remains a challenge for the scientific community. In many cases the quality of reconstructed electrical potential is not sufficiently accurate. The methods used consist in solving the Laplace equation on the volume delimited by the body surface and the epicardial surface.We plan to

- study in depth the dependance of this inverse problem inhomogeneities in the torso, conductivity values, the geometry, electrode placements...
- improve the solution to the inverse problem be using new regularization strategies and the theory of optimal control, both in the quasistatic and in the dynamic contexts.

Of cours we will use our models as a basis to regularize these inverse problems. We will conside the follwong strategies:

- using complete propagation models in the inverse problem, like the bidomain equations; for instance in order to localize some electrical sources;
- construct some families of reduced order models, using e.g. statistical learning techniques, which would accurately represent some families of well-identified pathologies;
- construct some simple models of the propagation of the activation front, based on eikonal or level-sets equations, but which would incorporate the representation of complex activation patterns.

Additionnaly, we will need to develop numerical techniques dedicated to our simplified eikonal/levl-sets equations.

3.3. Numerical techniques

We want the numerical simulations of the previous direct or inverse models to be efficient and reliable with respect to the need of the medical community. It needs to qualify and guarantee the accuracy and robustness of the numerical techniques and the efficiency of the resolution algorithms.

Based on previous work on solving the monodomain and bidomain equations [4], [5] and [6] and [1], we will focus on

- High-order numerical techniques with respect to the variables with physiological meaning, like velocity, AP duration and restitution properties;
- Efficient, dedicated preconditionning techniques coupled with parallel computing.

CASTOR Project-Team

3. Research Program

3.1. Plasma Physics

Participants: Jacques Blum, Cédric Boulbe, Blaise Faugeras, Hervé Guillard, Holger Heumann, Sebastian Minjeaud, Boniface Nkonga, Richard Pasquetti, Afeintou Sangam, Giorgio Giorgiani.

In order to fulfil the increasing demand, alternative energy sources have to be developed. Indeed, the current rate of fossil fuel usage and its serious adverse environmental impacts (pollution, greenhouse gas emissions, ...) lead to an energy crisis accompanied by potentially disastrous global climate changes.
Controlled fusion power is one of the most promising alternatives to the use of fossil resources, potentially with a unlimited source of fuel. France with the ITER (http://www.iter.org/default.aspx) and Laser Megajoule (http://www-lmj.cea.fr/) facilities is strongly involved in the development of these two parallel approaches to master fusion that are magnetic and inertial confinement. Although the principles of fusion reaction are well understood from nearly sixty years, (the design of tokamak dates back from studies done in the '50 by Igor Tamm and Andreï Sakharov in the former Soviet Union), the route to an industrial reactor is still long and the application of controlled fusion for energy production is beyond our present knowledge of related physical processes. In magnetic confinement, beside technological constraints involving for instance the design of plasma-facing component, one of the main difficulties in the building of a controlled fusion reactor is the poor confinement time reached so far. This confinement time is actually governed by turbulent transport that therefore determines the performance of fusion plasmas. The prediction of the level of turbulent transport in large machines such as ITER is therefore of paramount importance for the success of the researches on controlled magnetic fusion.
The other route for fusion plasma is inertial confinement. In this latter case, large scale hydrodynamical instabilities prevent a sufficient large energy deposit and lower the return of the target. Therefore, for both magnetic and inertial confinement technologies, the success of the projects is deeply linked to the theoretical understanding of plasma turbulence and flow instabilities as well as to mathematical and numerical improvements enabling the development of predictive simulation tools.

3.2. Turbulence Modelling

Participants: Boniface Nkonga, Richard Pasquetti.

Fluid turbulence has a paradoxical situation in science. The Navier-Stokes equations are an almost perfect model that can be applied to any flow. However, they cannot be solved for any flow of direct practical interest. Turbulent flows involve instability and strong dependence to parameters, chaotic succession of more or less organised phenomena, small and large scales interacting in a complex manner. It is generally necessary to find a compromise between neglecting a huge number of small events and predicting more or less accurately some larger events and trends.

In this direction, CASTOR wishes to contribute to the progress of methods for the prediction of fluid turbulence. Taking benefit of its experience in numerical methods for complex applications, CASTOR works out models for predicting flows around complex obstacles, that can be moved or deformed by the flow, and involving large turbulent structures. Taking into account our ambition to provide also short term methods for industrial problems, we consider methods applying to high Reynolds flows, and in particular, methods hybridizing Large Eddy Simulation (LES) with Reynolds Averaging.

Turbulence is the indirect cause of many other phenomena. Fluid-structure interaction is one of them, and can manifest itself for example in Vortex Induced Motion or Vibration. These phenomena can couple also with liquid-gas interfaces and bring new problems. Of particular interest is also the study of turbulence generated noise. In this field, though acoustic phenomena can also in principle be described by the Navier-Stokes equations, they are not generally numerically solved by flow solvers but rather by specialized linear and nonlinear acoustic solvers. An important question is the investigation of the best way to combine a LES simulation with the acoustic propagation of the waves it produces.

3.3. Astrophysical and Environmental flows

Participants: Didier Auroux, Hervé Guillard, Boniface Nkonga, Sebastian Minjeaud.

Although it seems inappropriate to address the modeling of experimental devices of the size of a tokamak and for instance, astrophysical systems with the same mathematical and numerical tools, it has long been recognized that the behaviour of these systems have a profound unity. This has for consequence for instance that any large conference on plasma physics includes sessions on astrophysical plasmas as well as sessions on laboratory plasmas. CASTOR does not intend to consider fluid models coming from Astrophysics or Environmental flows for themselves. However, the team is interested in the numerical approximation of some problems in this area as they provide interesting reduced models for more complex phenomena. To be more precise, let us give some concrete examples : The development of Rossby waves [0] a common problem in weather prediction has a counterpart in the development of magnetic shear induced instabilities in tokamaks and the understanding of this latter type of instabilities has been largely improved by the Rossby wave model. A second example is the water bag model of plasma physics that has a lot in common with multi-layer shallow water system.
To give a last example, we can stress that the development of the so-called well-balanced finite volume schemes used nowadays in many domains of mathematical physics or engineering was largely motivated by the desire to suppress some problems appearing in the approximation of the shallow water system.
Our goal is therefore to use astrophysical or geophysical models to investigate some numerical questions in contexts that, in contrast with plasma physics or fluid turbulence, do not require huge three dimensional computations but are still of interest for themselves and not only as toy models.

[0] Rossby waves are giant meanders in high altitude wind that have major influence on weather. Oceanic Rossby waves are also know to exist and to affect the world ocean circulation

CLIME Project-Team

3. Research Program

3.1. Data assimilation and inverse modeling

This activity is one major concern of environmental sciences. It matches up the setting and the use of data assimilation methods, for instance variational methods (such as the 4D-Var method). An emerging issue lies in the propagation of uncertainties by models, notably through ensemble forecasting methods.

Although modeling is not part of the scientific objectives of Clime, the project-team has complete access to models through collaborations with CEREA (Centre d'Enseignement et de Recherche en Environnement Atmosphérique, École des Ponts ParisTech): the models from Polyphemus (pollution forecasting from local to regional scales) and Code_Saturne (urban scale). In regard to other modeling domains, such as meteorology and oceanography, Clime accesses models through co-operation with LOCEAN (Laboratoire d'OCEANographie et du climat, UPMC).

The research activities of Clime tackle scientific issues such as:

- Within a family of models (differing by their physical formulations and numerical approximations), which is the optimal model for a given set of observations?
- How to reduce dimensionality of problems by Galerkin projection of equations on subspaces? How to define these subspaces in order to keep the main properties of systems?
- How to assess the quality of a forecast and its uncertainty? How do data quality, missing data, data obtained from sub-optimal locations, affect the forecast? How to better include information on uncertainties (of data, of models) within the data assimilation system?
- How to make a forecast (and a better forecast!) by using several models corresponding to different physical formulations? It also raises the question: how should data be assimilated in this context?
- Which observational network should be set up to perform a better forecast, while taking into account additional criteria such as observation cost? What are the optimal location, type and mode of deployment of sensors? How should trajectories of mobile sensors be operated, while the studied phenomenon is evolving in time? This issue is usually referred as "network design".

3.2. Satellite acquisitions and image assimilation

In geosciences, the issue of coupling data, in particular satellite acquisitions, and models is extensively studied for meteorology, oceanography, chemistry-transport and land surface models. However, satellite images are mostly assimilated on a point-wise basis. Three major approaches arise if taking into account the spatial structures, whose displacement is visualized on image sequences:

- Image approach. Image assimilation allows the extraction of features from image sequences, for instance motion field or structures' trajectory. A model of the dynamics is considered (obtained by simplification of a geophysical model such as Navier-Stokes equations). An observation operator is defined to express the links between the model state and the pixel values. In the simplest case, the pixel value corresponds to one coordinate of the model state and the observation operator is reduced to a projection. However, in most cases, this operator is highly complex, implicit and non-linear. Data assimilation techniques are developed to control the initial state or the whole assimilation window. Image assimilation is also applied to learn reduced models from image data and estimate a reliable and small-size reconstruction of the dynamics, which is observed on the sequence.
- Model approach. Image assimilation is used to control an environmental model and obtain improved forecasts. In order to take into account the spatial and temporal coherency of structures, specific image characteristics are considered and dedicated norms and observation error covariances are defined.

- Correcting a model. Another topic, mainly described for meteorology in the literature, concerns the location of structures. How to force the existence and to correct the location of structures in the model state using image information? Most of the operational meteorological forecasting institutes, such as Météo-France, UK-met, KNMI (in Netherlands), ZAMG (in Austria) and Met-No (in Norway), study this issue because operational forecasters often modify their forecasts based on visual comparisons between the model outputs and the structures displayed on satellite images.

3.3. Software chains for environmental applications

An objective of Clime is to participate in the design and creation of software chains for impact assessment and environmental crisis management. Such software chains bring together static or dynamic databases, data assimilation systems, forecast models, processing methods for environmental data and images, complex visualization tools, scientific workflows, ...

Clime is currently building, in partnership with École des Ponts ParisTech and EDF R&D, such a system for air pollution modeling: Polyphemus (see the web site http://cerea.enpc.fr/polyphemus/), whose architecture is specified to satisfy data requirements (e.g., various raw data natures and sources, data preprocessing) and to support different uses of an air quality model (e.g., forecasting, data assimilation, ensemble runs).

COFFEE Project-Team

3. Research Program

3.1. Research Program

Mathematical modeling and computer simulation are among the main research tools for environmental management, risks evaluation and sustainable development policy. Many aspects of the computer codes as well as the PDEs systems on which these codes are based can be considered as questionable regarding the established standards of applied mathematical modeling and numerical analysis. This is due to the intricate multiscale nature and tremendous complexity of those phenomena that require to set up new and appropriate tools. Our research group aims to contribute to bridging the gap by developing advanced abstract mathematical models as well as related computational techniques.

The scientific basis of the proposal is two–fold. On the one hand, the project is "technically–driven": it has a strong content of mathematical analysis and design of general methodology tools. On the other hand, the project is also "application–driven": we have identified a set of relevant problems motivated by environmental issues, which share, sometimes in a unexpected fashion, many common features. The proposal is precisely based on the conviction that these subjects can mutually cross-fertilize and that they will both be a source of general technical developments, and a relevant way to demonstrate the skills of the methods we wish to design.

To be more specific:

- We consider evolution problems describing highly heterogeneous flows (with different phases or with high density ratio). In turn, we are led to deal with non linear systems of PDEs of convection and/or convection–diffusion type.
- The nature of the coupling between the equations can be two–fold, which leads to different difficulties, both in terms of analysis and conception of numerical methods. For instance, the system can couple several equations of different types (elliptic/parabolic, parabolic/hyperbolic, parabolic or elliptic with algebraic constraints, parabolic with degenerate coefficients....). Furthermore, the unknowns can depend on different sets of variables, a typical example being the fluid/kinetic models for particulate flows. In turn, the simulation cannot use a single numerical approach to treat all the equations. Instead, hybrid methods have to be designed which raise the question of fitting them in an appropriate way, both in terms of consistency of the discretization and in terms of stability of the whole computation. For the problems under consideration, the coupling can also arises through interface conditions. It naturally occurs when the physical conditions are highly different in subdomains of the physical domain in which the flows takes place. Hence interface conditions are intended to describe the exchange (of mass, energy...) between the domains. Again it gives rise to rather unexplored mathematical questions, and for numerics it yields the question of defining a suitable matching at the discrete level, that is requested to preserve the properties of the continuous model.
- By nature the problems we wish to consider involve many different scales (of time or length basically). It raises two families of mathematical questions. In terms of numerical schemes, the multiscale feature induces the presence of stiff terms within the equations, which naturally leads to stability issues. A clear understanding of scale separation helps in designing efficient methods, based on suitable splitting techniques for instance. On the other hand asymptotic arguments can be used to derive hierarchy of models and to identify physical regimes in which a reduced set of equations can be used.

We can distinguish the following fields of expertise

- Numerical Analysis: Finite Volume Schemes, Well-Balanced and Asymptotic-Preserving Methods
  - Finite Volume Schemes for Diffusion Equations
  - Finite Volume Schemes for Conservation Laws
  - Well-Balanced and Asymptotic-Preserving Methods
- Modeling and Analysis of PDEs
  - Kinetic equations and hyperbolic systems
  - PDEs in random media
  - Interface problems

DEMAR Project-Team

3. Research Program

3.1. Modelling and identification of the sensory-motor system

Participants: Mitsuhiro Hayashibe, Christine Azevedo Coste, David Guiraud.

The literature on muscle modelling is vast, but most of research works focus separately on the microscopic and on the macroscopic muscle's functional behaviours. The most widely used microscopic model of muscle contraction was proposed by Huxley in 1957. The Hill-Maxwell macroscopic model was derived from the original model introduced by A.V. Hill in 1938. We may mention the most recent developments including Zahalak's work introducing the distribution moment model that represents a formal mathematical approximation at the sarcomere level of the Huxley cross-bridges model and the works by Bestel and Sorine (2001) who proposed an explanation of the beating of the cardiac muscle by a chemical control input connected to the calcium dynamics in the muscle cells, that stimulates the contractile elements of the model. With respect to this literature, our contributions are mostly linked with the model of the contractile element, through the introduction of the recruitment at the fibre scale formalizing the link between FES parameters, recruitment and Calcium signal path. The resulting controlled model is able to reproduce both short term (twitch) and long term (tetanus) responses. It also matches some of the main properties of the dynamic behaviour of muscles, such as the Hill force-velocity relationship or the instantaneous stiffness of the Mirsky-Parmley model. About integrated functions modelling such as spinal cord reflex loops or central pattern generator, much less groups work on this topic compared to the ones working on brain functions. Mainly neurophysiologists work on this subject and our originality is to combine physiology studies with mathematical modelling and experimental validation using our own neuroprostheses. The same analysis could be drawn with sensory feedback modelling. In this domain, our work is based on the recording and analysis of nerve activity through electro-neurography (ENG). We are interested in interpreting ENG in terms of muscle state in order to feedback useful information for FES controllers and to evaluate the stimulation effect. We believe that this knowledge should help to improve the design and programming of neuroprostheses. We investigate risky but promising fields such as intrafascicular recordings, area on which only few teams in North America (Canada and USA), and Denmark really work on. Very few teams in France, and none at Inria work on the peripheral nervous system modelling, together with experimental protocols that need neuroprostheses. Most of our Inria collaborators work on the central nervous system, except the spinal cord, (ODYSSEE for instance), or other biological functions (SISYPHE for instance). Our contributions concern the following aspects:

- Muscle modelling,
- Sensory organ modelling,
- Electrode nerve interface,
- High level motor function modelling,
- Model parameters identification.

We contribute both to the design of reliable and accurate experiments with a well-controlled environment, to the fitting and implementation of efficient computational methods derived for instance from Sigma Point Kalman Filtering.

3.2. Synthesis and Control of Human Functions

Participants: Christine Azevedo Coste, Philippe Fraisse, Mitsuhiro Hayashibe, David Andreu.

We aim at developping realistic solutions for real clinical problems expressed by patients and medical staff. Different approaches and specifications are developped to answer those issues in short, mid or long terms. This research axis is therefore obviously strongly related to clinical application objectives. Even though applications can appear very different, the problematic and constraints are usually similar in the context of electrical stimulation: classical desired trajectory tracking is not possible, robustness to disturbances is critical, possible observations of system are limited. Furthermore there is an interaction between body segments under voluntary control of the patient and body segments under artificial control. Finally, this axis relies on modelling and identification results obtained in the first axis and on the technological solutions and approaches developped in the third axis (Neuroprostheses). The robotics framework involved in DEMAR work is close to the tools used and developped by BIPOP team in the context of bipedal robotics. There is no national team working on those aspects. Within international community, several colleagues carry out researches on the synthesis and control of human functions, most of them belong to the International Functional Electrical Stimulation Society (IFESS) community. In the following we present two sub-objectives. Concerning spinal cord injuries (SCI) context not so many team are now involved in such researches around the world. Our force is to have technological solutions adapted to our theoretical developments. Concerning post-stroke context, several teams in Europe and North America are involved in drop-foot correction using FES. Our team specificity is to have access to the different expertises needed to develop new theoretical and technical solutions: medical expertise, experimental facilities, automatic control expertise, technological developments, industrial partner. These expertises are available in the team and through strong external collaborations.

3.3. Neuroprostheses

Participants: David Andreu, David Guiraud, Daniel Simon, Guy Cathébras, Fabien Soulier.

The main drawbacks of existing implanted FES systems are well known and include insufficient reliability, the complexity of the surgery, limited stimulation selectivity and efficiency, the non-physiological recruitment of motor units and muscle control. In order to develop viable implanted neuroprostheses as palliative solutions for motor control disabilities, the third axis "Neuroprostheses" of our project-team aims at tackling four main challenges: (i) a more physiologically based approach to muscle activation and control, (ii) a fibres' type and localization selective technique and associated technology (iii) a neural prosthesis allowing to make use of automatic control theory and consequently real-time control of stimulation parameters, and (iv) small, reliable, safe and easy-to-implant devices.

Accurate neural stimulation supposes the ability to discriminate fibres' type and localization in nerve and propagation pathway; we thus jointly considered multipolar electrode geometry, complex stimulation profile generation and neuroprosthesis architecture. To face stimulation selectivity issues, the analog output stage of our stimulus generator responds to the following specifications: i) temporal controllability in order to generate current shapes allowing fibres' type and propagation pathway selectivity, ii) spatial controllability of the current applied through multipolar cuff electrodes for fibres' recruitment purposes. We have therefore proposed and patented an original architecture of output current splitter between active poles of a multipolar electrode. The output stage also includes a monotonic DAC (Digital to Analog Converter) by design. However, multipolar electrodes lead to an increasing number of wires between the stimulus generator and the electrode contacts (poles); several research laboratories have proposed complex and selective stimulation strategies involving multipolar electrodes, but they cannot be implanted if we consider multisite stimulation (i.e. stimulating on several nerves to perform a human function as a standing for instance). In contrast, all the solutions tested on humans have been based on centralized implants from which the wires output to only monopolar or bipolar electrodes, since multipolar ones induce to many wires. The only solution is to consider a distributed FES architecture based on communicating controllable implants. Two projects can be cited: Bion technology (main competitor to date), where bipolar stimulation is provided by injectable autonomous units, and the LARSI project, which aimed at multipolar stimulation localized to the sacral roots. In both cases, there was no application breakthrough for reliable standing or walking for paraplegics. The power source, square stimulation shape and bipolar electrode limited the Bion technology, whereas the insufficient selection accuracy of the LARSI implant disqualified it from reliable use.

Keeping the electronics close to the electrode appears to be a good, if not the unique, solution for a complex FES system; this is the concept according to which we direct our neuroprosthesis design and development, in close relationship with other objectives of our project-team (control for instance) but also in close collaboration with medical and industrial partners.
Our efforts are mainly directed to implanted FES systems but we also work on surface FES architecture and stimulator; most of our concepts and advancements in implantable neuroprostheses are applicable somehow to external devices.

DRACULA Project-Team

3. Research Program

3.1. Cell dynamics

We model dynamics of cell populations with two approaches, dissipative particle dynamics (DPD) and partial differential equations (PDE) of continuum mechanics. DPD is a relatively new method developed from molecular dynamics approach largely used in statistical physics. Particles in DPD do not necessarily correspond to atoms or molecules as in molecular dynamics. These can be mesoscopic particles. Thus, we describe in this approach a system of particles. In the simplest case where each particle is a sphere, they are characterized by their positions and velocities. The motion of particles is determined by Newton's second law (see Figure 1).

In our case, particles correspond to biological cells. The specific feature of this case in comparison with the conventional DPD is that cells can divide (proliferation), change their type (differentiation) and die by apoptosis or necrosis. Moreover, they interact with each other and with the extra-cellular matrix not only mechanically but also chemically. They can exchange signals, they can be influenced by various substances (growth factors, hormones, nutrients) coming from the extra-cellular matrix and, eventually, from other organs.

Distribution of the concentrations of bio-chemical substances in the extra-cellular matrix will be described by the diffusion equation with or without convective terms and with source and/or sink terms describing their production or consumption by cells. Thus we arrive to a coupled DPD-PDE model.

Cell behaviour (proliferation, differentiation, apoptosis) is determined by intra-cellular regulatory networks, which can be influenced by external signals. Intra-cellular regulatory networks (proteins controlling the cell cycle) can be described by systems of ordinary differential equations (ODE). Hence we obtain DPD-PDE-ODE models describing different levels of cell dynamics (see Figure 1). It is important to emphasize that the ODE systems are associated to each cell and they can depend on the cell environment (extra-cellular matrix and surrounding cells).

3.2. From particle dynamics to continuum mechanics

DPD is well adapted to describe biological cells. However, it is a very time consuming method which becomes difficult to use if the number of particles exceeds the order of 10^5-10^6 (unless distributed computing is used). On the other hand, PDEs of continuum mechanics are essentially more efficient for numerical simulations. Moreover, they can be studied by analytical methods which have a crucial importance for the understanding of relatively simple test cases. Thus we need to address the question about the relation between DPD and PDE. The difficulty follows already from the fact that molecular dynamics with the Lennard-Jones potential can describe very different media, including fluids (compressible, incompressible, non-Newtonian, and so on) and solids (elastic, elasto-plastic, and so on). Introduction of dissipative terms in the DPD models can help to justify the transition to a continuous medium because each medium has a specific to it law of dissipation. Our first results [39] show the correspondence between a DPD model and Darcy's law describing fluid motion in a porous medium. However, we cannot expect a rigorous justification in the general case and we will have to carry out numerical comparison of the two approaches.

An interesting approach is related to hybrid models where PDEs of continuum mechanics are considered in the most part of the domain, where we do not need a microscopical description, while DPD in some particular regions are required to consider individual cells.

3.3. PDE models

If we consider cell populations as a continuous medium, then cell concentrations can be described by reaction-diffusion systems of equations with convective terms. The diffusion terms correspond to a random cell motion and the reaction terms to cell proliferation, differentiation and death. These are more traditional models [40] with properties that depend on the particular problem under consideration and with many open questions, both from the point of view of their mathematical properties and for applications. In particular we are interested in the spreading of cell populations which describes the development of leukemia in the bone marrow and many other biological phenomena (solid tumors, morphogenesis, atherosclerosis, and so on). From the mathematical point of view, these are reaction-diffusion waves, intensively studied in relation with various biological problems. We will continue our studies of wave speed, stability, nonlinear dynamics and pattern formation. From the mathematical point of view, these are elliptic and parabolic problems in bounded or unbounded domains, and integro-differential equations. We will investigate the properties of the corresponding linear and nonlinear operators (Fredholm property, solvability conditions, spectrum, and so on). Theoretical investigations of reaction-diffusion-convection models will be accompanied by numerical simulations and will be applied to study hematopoiesis.

Hyperbolic problems are also of importance when describing cell population dynamics ([45], [47]), and they proved effective in hematopoiesis modelling ([34], [35], [37]). They are structured transport partial differential equations, in which the structure is a characteristic of the considered population, for instance age, size, maturity, protein concentration, etc. The transport, or movement in the structure space, simulates the progression of the structure variable, growth, maturation, protein synthesis, etc. Several questions are still open in the study of transport PDE, yet we will continue our analysis of these equations by focusing in particular on the asymptotic behaviour of the system (stability, bifurcation, oscillations) and numerical simulations of nonlocal transport PDE.

The use of age structure often leads to a reduction (by integration over the age variable) to nonlocal problems [47]. The nonlocality can be either in the structure variable or in the time variable [34]. In particular, when coefficients of an age-structured PDE are not supposed to depend on the age variable, this reduction leads to delay differential equations.

3.4. Delay differential Equations

Delay differential equations (DDEs) are particularly useful for situations where the processes are controlled through feedback loops acting after a certain time. For example, in the evolution of cell populations the transmission of control signals can be related to some processes as division, differentiation, maturation, apoptosis, etc. Because these processes can take a certain time, the system depends on an essential way of its past state, and can be modelled by DDEs.

We explain hereafter how delays can appear in hematopoietic models. Based on biological aspects, we can divide hematopoietic cell populations into many compartments. We basically consider two different cell populations, one composed with immature cells, and the other one made of mature cells. Immature cells are separated in many stages (primitive stem cells, progenitors and precursors, for example) and each stage is composed with two sub-populations, resting (G0) and proliferating cells. On the opposite, mature cells are known to proliferate without going into the resting compartment. Usually, to describe the dynamic of these multi-compartment cell populations, transport equations (hyperbolic PDEs) are used. Structure variables are age and discrete maturity. In each proliferating compartment, cell count is controlled by apoptosis (programmed cell death), and in the other compartments, cells can be eliminated only by necrosis (accidental cell death). Transitions between the compartments are modelled through boundary conditions. In order to reduce the complexity of the system and due to some lack of information, no dependence of the coefficients on cell age is assumed. Hence, the system can be integrated over the age variable and thus, by using the method of characteristics and the boundary conditions, the model reduces to a system of DDEs, with several delays.

Leaving all continuous structures, DDEs appear well adapted to us to describe the dynamics of cell populations. They offer good tools to study the behaviour of the systems. The main investigation of DDEs are the effect of perturbations of the parameters, as cell cycle duration, apoptosis, differentiation, self-renewal, and re-introduction from quiescent to proliferating phase, on the behaviour of the system, in relation for instance with some hematological disorders [41].

DYLISS Project-Team

3. Research Program

3.1. Knowledge representation with constraint programming

Biological networks are built with data-driven approaches aiming at translating genomic information into a functional map. Most methods are based on a probabilistic framework which defines a probability distribution over the set of models. The reconstructed network is then defined as the most likely model given the data. In the last few years, our team has investigated an alternative perspective where each observation induces a set of constraints - related to the steady state response of the system dynamics - on the set of possible values in a network of fixed topology. The methods that we have developed complete the network with product states at the level of nodes and influence types at the level of edges, able to globally explain experimental data. In other words, the selection of relevant information in the model is no more performed by selecting *the* network with the highest score, but rather by exploring the complete space of models satisfying constraints on the possible dynamics supported by prior knowledge and observations. In the (common) case when there is no model satisfying all the constraints, we need to relax the problem and to study the space of corrections to prior knowledge in order to fit reasonably with observation data. In this case, this issue is modeled as combinatorial (sub)-optimization issues. In both cases, common properties to all solutions are considered as a robust information about the system, as they are independent from the choice of a single solution to the satisfiability problem (in the case of existing solutions) or to the optimization problem (in the case of required corrections to the prior knowledge) [5].

Solving these computational issues requires addressing NP-hard qualitative (non-temporal) issues. We have developed a long-term collaboration with Potsdam University in order to use a logical paradigm named **Answer Set Programming** [45], [53] to solve these constraint satisfiability and combinatorial optimization issues. Applied on transcriptomic or cancer networks, our methods identified which regions of a large-scale network shall be corrected [46], and proposed robust corrections [4]. See Fig. 1 for details. The results obtained so far suggest that this approach is compatible with efficiency, scale and expressivity needed by biological systems. Our goal is now to provide **formal models of queries on biological networks** with the focus of integrating dynamical information as explicit logical constraints in the modeling process. This would definitely introduce such logical paradigms as a powerful approach to build and query reconstructed biological systems, in complement to discriminative approaches. Notice that our main issue is in the field of knowledge representation. More precisely, we do not wish to develop new solvers or grounders, a self-contained computational issue which is addressed by specialized teams such as our collaborator team in Potsdam. Our goal is rather to investigate whether progresses in the field of constraint logical programming, shown by the performance of ASP-solvers in several recent competitions, are now sufficient to address the complexity of constraint-satisfiability and combinatorial optimization issues explored in systems biology.

By exploring the complete space of models, our approach typically produces numerous candidate models compatible with the observations. We began investigating to what extent domain knowledge can further refine the analysis of the set of models by identifying classes of similar models, or by selecting the models that best fit biological knowledge. We anticipate that this will be particularly relevant when studying non-model species for which little is known but valuable information from other species can be transposed or adapted. These efforts consist in developing reasoning methods based on ontologies as formal representation of symbolic knowledge. We use Semantic Web tools such as SPARQL for querying and integrating large sources of external knowledge, and measures of semantic similarity and particularity for analyzing data.

Using these technologies requires to revisit and reformulate constraint-satisfiability problems at hand in order both to decrease the search space size in the grounding part of the process and to improve the exploration of this search space in the solving part of the process. Concretely, getting logical encoding for the optimization problems forces to clarify the roles and dependencies between parameters involved in the problem. This opens

the way to a refinement approach based on a fine investigation of the space of hypotheses in order to make it smaller and gain in the understanding of the system.

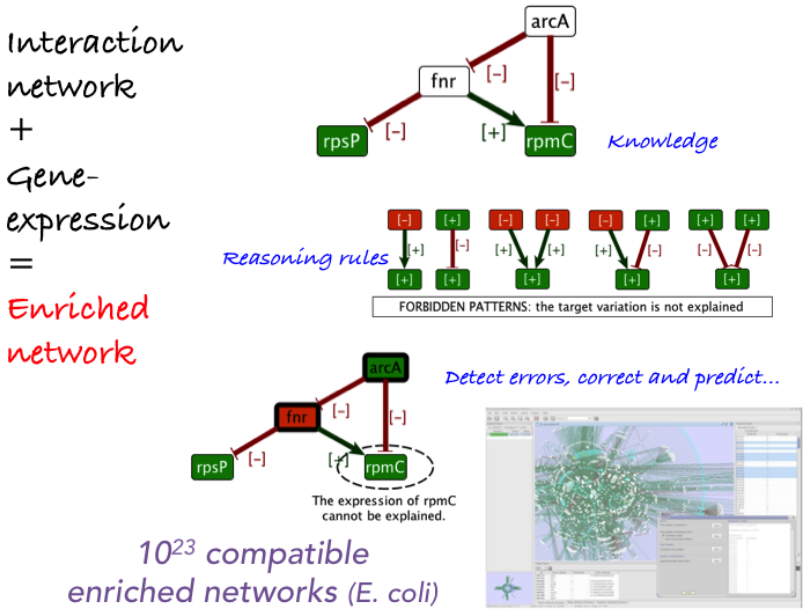


Figure 1.

An example of reasoning process in order to identify which expression of non-observed nodes (white nodes) are fixed by partial observations and rules derived from the system dynamics. The ASP-based logical approach is flexible enough to model in a single framework network characteristics (products, interactions, partial information on signs of regulations and observations) and static rules about the effects of the dynamics of the system. Extensions of this framework include the exhaustive search for system repair or more constrained dynamical rules. [5], [4] ***Step 1****. Regulation knowledge is represented as a signed oriented graph. Edge colors stand for regulatory effects (red/green→ inhibition or activation). Vertex colors stand for gene expression data (red/green→ under or over-expression).* ***Step 2****. Integrity constraints on the whole colored graph come from the necessity to find a consistent explanation of the link between regulation and expression.* ***Step 3****. The model allows both the prediction of values (e.g. for fnr in the figure) and the detection of contradictions (e.g. the expression level of rpmC is inconsistant with the regulation in the graph). .*

3.2. Probabilistic and symbolic dynamics

We work on new techniques to emphasize biological strategies that must occur to reproduce quantitative measurements in order to predict the quantitative response of a system at a larger-scale. Our framework mixes mechanistic and probabilistic modeling [1]. The system is modeled by an Event Transition Graph, that is, a **Markovian qualitative description of its dynamics** together with quantitative laws which describe the effect of the dynamic transitions over higher scale quantitative measurements. Then, a few time-series quantitative measurements are provided. Following an ergodic assumption and average case analysis properties, we know that a multiplicative accumulation law on a Markov chain asymptotically follows a log-normal law with explicit parameters [52]. This property can be derived into constraints to describe the set of admissible weighted Markov chains whose asymptotic behavior agrees with the quantitative measures at hand. A precise

study of this constrained space via local search optimization emphasizes the most important discrete events that must occur to reproduce the information at hand. These methods have been validated on the *E. coli* regulatory network benchmark. See Figure 2 for illustration. We now plan to apply these techniques to reduced networks representing the main pathways and actors automatically generated from the integrative methods developed in the former section. This requires to improve the range of dynamics that can be modeled by these techniques, as well as the efficiency and scalability of the local search algorithms.

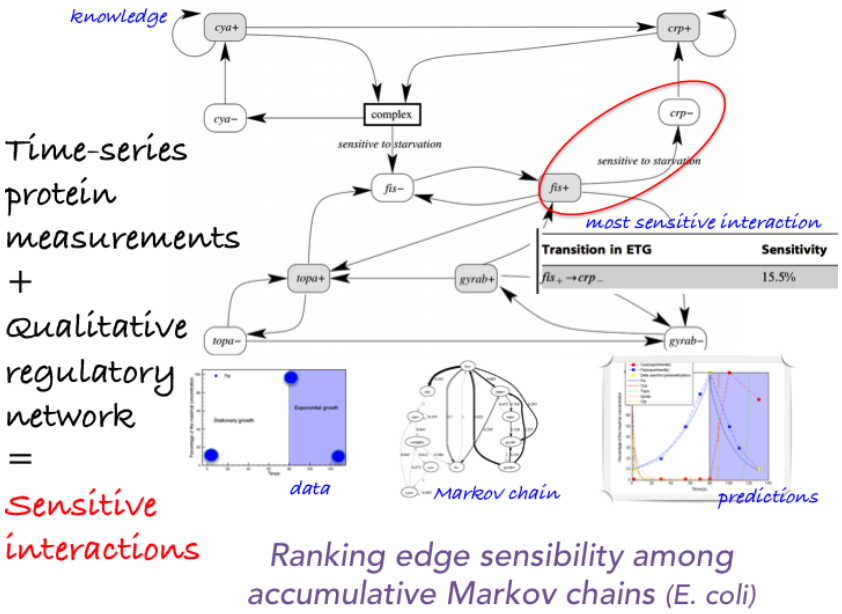


Figure 2.

Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems and identification of key interactions [1].
Input data *are provided by a qualitative description of the system dynamics at the transcription level (interaction graph) and 3 concentration measurements of the* fis *protein (population scale). The method computes an* ***Event-Transition Graph****. Interaction frequencies required to predict the population scale behavior as the asymptotic behavior of an accumulation multiplicative law over a Markov chain. Estimation by local searches in the space of Markov chains consistent with the observed dynamics and whose asymptotic behavior is consistent with quantitative observations at the population scale. Edge thickness reflects their sensitivity in the search space. It allows to* ***predict*** *the* Cya *protein concentration (red curve) which fits with observations. Additionally, literature evidences that high sensitivity ETG transitions correspond to key interaction in E. Coli response to nutritional stress.*

3.3. Grammatical inference and highly expressive structures

Our main field of expertise in machine learning concerns grammatical models with a strong expertise in finite state automata learning. By introducing a similar fragment merging heuristic approach, we have proposed an algorithm that learns successfully automata modeling families of (non homologous) functional families of proteins [3], leading to a tool named Protomata-learner. As an example, this tools allows us to properly model the function of the protein family TNF, which is impossible with other existing probabilistic-based approach (see Fig. 3). It was also applied to model families of proteins in cyanobacteria [2]. Our future goal

is to further demonstrate the relevance of formal language modelling by addressing the question of enzyme prediction, from their genomic or protein sequences, aiming at better sensitivity and specificity. As enzyme-substrate interactions are very specific central relations for integrated genome/metabolome studies and are characterized by faint signatures, we shall rely on models for active sites involved in cellular regulation or catalysis mechanisms. This requires to build models gathering both structural and sequence information is order to describe (potentially nested or crossing) long-term dependencies such as contacts of amino-acids that are far in the sequence but close in the 3D protein folding. We wish to extend our expertise towards inferring Context-Free Grammars including the topological information coming from the structural characterization of active sites.

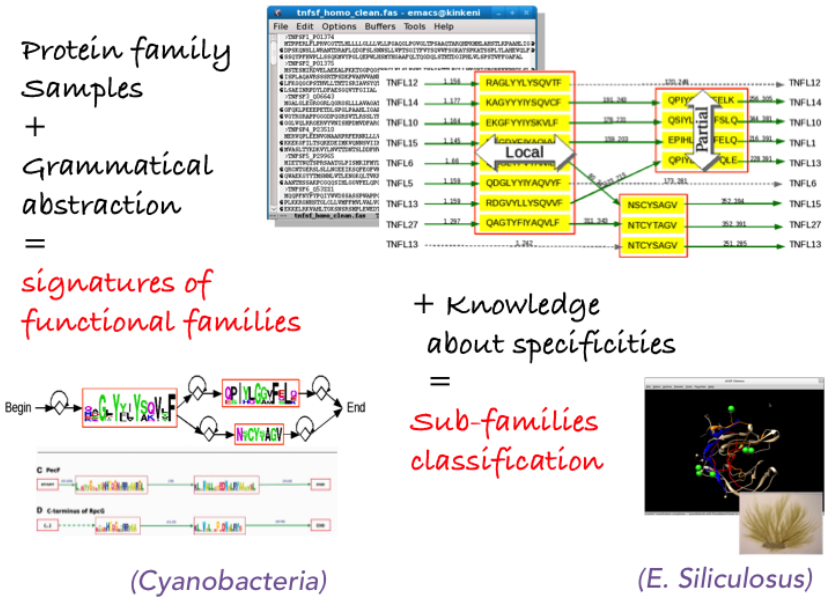
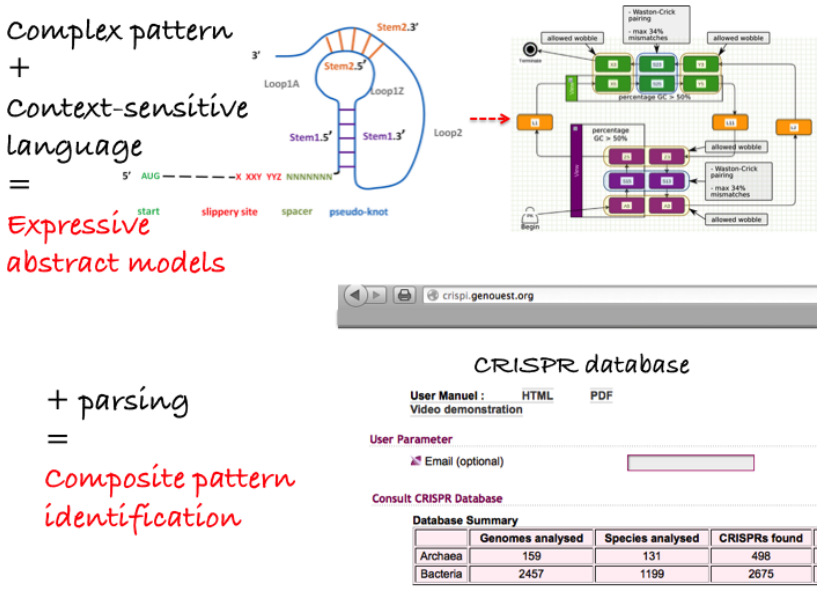


Figure 3. ***Protomata Learner workflow****. Starting from a set of protein sequences, a partial local alignment is computed and an automaton is inferred, which can be considered as a signature of the family of proteins. This allows searching for new members of the family [2]. Adding further information about the specific properties of proteins within the family allows to exhibit a refined classification.*

Using context-free grammars instead of regular patterns increases the complexity of parsing issues. Indeed, efficient parsing tools have been developed to identify patterns within genomes but most of them are restricted to simple regular patterns. Definite Clause Grammars (DCG), a particular form of logical context-free grammars have been used in various works to model DNA sequence features [54]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [59], [58]. This increases the expressivity of the formalism towards midly context sensitive grammars. Thus, those grammars model not only DNA/RNA sequence features but also structural features such as repeats, palindromes, stem/loop or pseudo-knots. We have designed a tool, STAN (suffix-tree analyser) which makes it possible to search for a subset of SVG patterns in full chromosome sequences [7]. This tool was used for the recognition of transposable elements in *Arabidopsis thaliana* [60] or for the design of a CRISPR database [9]. See Figure 4 for illustration. Our goal is to extend the framework of STAN. Generally, a suitable language for the search of particular components in languages has to meet several needs : expressing existing structures in a compact way, using existing databases of motifs, helping the description of interacting

components. In other words, the difficulty is to find a good tradeoff between expressivity and complexity to allow the specification of realistic models at genome scale. In this direction, we are working on Logol, a language and framework based on a systematic introduction of constraints on string variables.



	Genomes analysed	Species analysed	CRISPRs found
Archaea	159	131	498
Bacteria	2457	1199	2675

Figure 4. Graphical modeling of a pseudo-knot (RNA structure) based on the expressivity of String Variable Grammars used in the Logol framework. Combined with parsers, this leads to composite pattern identification such as CRISPR [56].

FLUMINANCE Project-Team

3. Research Program

3.1. Estimation of fluid characteristic features from images

The measurement of fluid representative features such as vector fields, potential functions or vorticity maps, enables physicists to have better understanding of experimental or geophysical fluid flows. Such measurements date back to one century and more but became an intensive subject of research since the emergence of correlation techniques [29] to track fluid movements in pairs of images of a particles laden fluid or by the way of clouds photometric pattern identification in meteorological images. In computer vision, the estimation of the projection of the apparent motion of a 3D scene onto the image plane, referred to in the literature as optical-flow, is an intensive subject of researches since the 80's and the seminal work of B. Horn and B. Schunk [42]. Unlike to dense optical flow estimators, the former approach provides techniques that supply only sparse velocity fields. These methods have demonstrated to be robust and to provide accurate measurements for flows seeded with particles. These restrictions and their inherent discrete local nature limit too much their use and prevent any evolutions of these techniques towards the devising of methods supplying physically consistent results and small scale velocity measurements. It does not authorize also the use of scalar images exploited in numerous situations to visualize flows (image showing the diffusion of a scalar such as dye, pollutant, light index refraction, flurocein,...). At the opposite, variational techniques enable in a well-established mathematical framework to estimate spatially continuous velocity fields, which should allow more properly to go towards the measurement of smaller motion scales. As these methods are defined through PDE's systems they allow quite naturally including as constraints kinematic properties or dynamic laws governing the observed fluid flows. Besides, within this framework it is also much easier to define characteristic features estimation procedures on the basis of physically grounded data model that describes the relation linking the observed luminance function and some state variables of the observed flow.

A substantial progress has been done in this direction with the design of dedicated dense estimation techniques to estimate dense fluid motion fields[4], [10], the setting up of tomographic techniques to carry out 3D velocity measurements [36], the inclusion of physical constraints to infer 3D motions or the design of dynamically consistent velocity measurements to provide coherent motion fields from time resolved fluid flow image sequences [9]. These progresses have brought further accuracy and an improved spatial resolution for a variety of applications ranging from experimental fluid mechanics to geophysical sciences. For a detailed review of these approaches see [7].

We believe that such approaches must be first enlarged to the wide variety of imaging modalities enabling the observation of fluid flows. This covers for instance, the systematic study of motion estimation for the different channels of meteorological satellites, but also of other experimental imaging tools such as Shadowgraphs, Background oriented Schlieren, Schlieren [49], diffusive scalar images, fluid holography [50], or Laser Induced Fluorimetry. All these modalities offer the possibility to visualize time resolved sequences of the flow. The velocity measurement processes available to date for that kind of images suffer from a lack of physical relevancy to keep up with the increasing amount of fine and coherent information provided by the images. We think, and have begun to prove, that a significant step forward can be taken by providing new tools based on sound data models and adapted regularization functional, both built on physical grounds.

Additional difficulties arise when considering the necessity to go towards 3D measurements and 3D volumetric reconstruction of the observed flows (e.g., the tomographic PIV paradigm). First, unlike in the standard setup, the 2D images captured by the experimentalists only provide a partial information about the structure of the particles transported by the fluid. As a matter of fact, inverse problems have to be solved in order to recover this crucial information. Secondly, another issue stands in the increase of the underdetermination of the problem, that is the important decrease of the ratio between the number of observations and the total number of unknowns. In particular, this point asks for methodologies able to gather and exploit observations

captured at different time instants. Finally, the dimensions of the problem (that is, the number of unknown) dramatically increase with the transition from the 2D to the 3D paradigm. This leads, as a by-product, to a significant amplification of the computational burden and requires the conception of efficient algorithms, exhibiting a reasonable scaling with the problem dimensions.

The first problem can be addressed by resorting to state-of-the-art methodologies pertaining to sparse representations. These techniques consist in identifying the solution of an inverse problem with the most “zero" components which, in the case of the tomographic PIV, turns out to be a physically relevant option. Hence, the design of sparse representation algorithms and the study of their conditions of success constitute an important research topic of the group. On the other hand, we believe that the dramatic increase of the under-determination appearing in the 3D setup can be tackled by combining tomographic reconstruction of several planar views of the flow with data assimilation techniques. These techniques enable to couple a dynamical model with incomplete observations of the flow. Each applicative situation under concern defines its proper required scale of measurement and a scale for the dynamical model. For instance, for control or monitoring purposes, very rapid techniques are needed whereas for analysis purpose the priority is to get accurate measurements of the smallest motion scales as possible. These two extreme cases imply the use of different models but also of different algorithmic techniques. Recursive techniques and large scale representation of the flow are relevant for the first case whereas batch techniques relying on the whole set of data available and models refined down to small scales have to be used for the latter case.

The question of the scale of the velocity measurement is also an open question that must be studied carefully. Actually, no scale considerations are taken into account in the estimation schemes. It is more or less abusively assumed that the measurements supplied have a subpixel accuracy, which is obviously erroneous due to implicit smoothness assumptions made either in correlation techniques or in variational estimation techniques. We are convinced that to go towards the measurement of the smaller scales of the flow it is necessary to introduce some turbulence or uncertainty subgrid modeling within the estimation scheme and also to devise alternative regularization schemes that fit well with phenomenological statistical descriptions of turbulence described by the velocity increments moments. As a by product such schemes should offer the possibility to have a direct characterization, from image sequences, of the flow turbulent regions in term of vortex tube, area of pure straining, or vortex sheet. This philosophy should allow us to elaborate methods enabling the estimation of relevant characteristics of the turbulence like second-order structure functions, mean energy dissipation rate, turbulent viscosity coefficient, or dissipative scales.

We are planning to study these questions for a wide variety of application domains ranging from experimental fluid mechanics to geophysical sciences. We believe there are specific needs in different application domains that require clearly identified developments and modeling. Let us for instance mention meteorology and oceanography which both involve very specific dynamical modeling but also micro-fluidic applications or bio-fluid applications that are ruled by other types of dynamics.

3.2. Data assimilation and Tracking of characteristic fluid features

Real flows have an extent of complexity, even in carefully controlled experimental conditions, which prevents any set of sensors from providing enough information to describe them completely. Even with the highest levels of accuracy, space-time coverage and grid refinement, there will always remain at least a lack of resolution and some missing input about the actual boundary conditions. This is obviously true for the complex flows encountered in industrial and natural conditions, but remains also an obstacle even for standard academic flows thoroughly investigated in research conditions.

This unavoidable deficiency of the experimental techniques is nevertheless more and more compensated by numerical simulations. The parallel advances in sensors, acquisition, treatment and computer efficiency allow the mixing of experimental and simulated data produced at compatible scales in space and time. The inclusion of dynamical models as constraints of the data analysis process brings a guaranty of coherency based on fundamental equations known to correctly represent the dynamics of the flow (e.g. Navier Stokes equations) [3], [5].

Conversely, the injection of experimental data into simulations ensures some fitting of the model with reality. When used with the correct level of expertise to calibrate the models at the relevant scales, regarding data validity and the targeted representation scale, this collaboration represents a powerful tool for the analysis and reconstruction of the flows. Automated back and forth sequencing between data integration and calculations have to be elaborated for the different types of flows with a correct adjustment of the observed and modeled scales. This appears more and more feasible when considering the sensitivity, the space resolution and above all the time resolution that the imaging sensors are reaching now.

That becomes particularly true, for instance, for satellite imaging, the foreseeable advances of which will soon give the right complement to the progresses in atmospheric and ocean modeling to dramatically improve the analysis and predictions of physical states and streams for weather and environment monitoring. In that domain, there is a particular interest in being able to combine image data, models and in-situ measurements, as high densities of data supplied by meteorological stations are available only for limited regions of the world, typically Europe and USA, while Africa, or the south hemisphere lack of refined and frequent *in situ* measurements. Moreover, we believe that such an approach can favor great advances in the analysis and prediction of complex flows interactions like those encountered in sea-atmosphere interactions, dispersion of polluting agents in seas and rivers, etc. In other domains we believe that image data and dynamical models coupling may bring interesting solutions for the analysis of complex phenomena which involve multi-phasic flows, interaction between fluid and structures, and the general case of flows with complex unknown border conditions.

The coupling approach can be extended outside the fluidics domain to complex dynamics that can be modeled either from physical laws or from learning strategies based on the observation of previous events [1]. This concerns for instance forest combustion, the analysis of the biosphere evolution, the observation and prediction of the melting of pack ice, the evolution of sea ice, the study of the consequences of human activity like deforestation, city growing, landscape and farming evolution, etc. All these phenomena are nowadays rapidly evolving due to global warming. The measurement of their evolution is a major societal interest for analysis purpose or risk monitoring and prevention.

To enable data and models coupling to achieve its potential, some difficulties have to be tackled. It is in particular important to outline the fact that the coupling of dynamical models and image data are far from being straightforward. The first difficulty is related to the space of the physical model. As a matter of fact, physical models describe generally the phenomenon evolution in a 3D Cartesian space whereas images provides generally only 2D tomographic views or projections of the 3D space on the 2D image plane. Furthermore, these views are sometimes incomplete because of partial occlusions and the relations between the model state variables and the image intensity function are otherwise often intricate and only partially known. Besides, the dynamical model and the image data may be related to spatio-temporal scale spaces of very different natures which increases the complexity of an eventual multiscale coupling. As a consequence of these difficulties, it is necessary generally to define simpler dynamical models in order to assimilate image data. This redefinition can be done for instance on an uncertainty analysis basis, through physical considerations or by the way of data based empirical specifications. Such modeling comes to define inexact evolution laws and leads to the handling of stochastic dynamical models. The necessity to make use and define sound approximate models, the dimension of the state variables of interest and the complex relations linking the state variables and the intensity function, together with the potential applications described earlier constitute very stimulating issues for the design of efficient data-model coupling techniques based on image sequences.

On top of the problems mentioned above, the models exploited in assimilation techniques often suffer from some uncertainties on the parameters which define them. Hence, a new emerging field of research focuses on the characterization of the set of achievable solutions as a function of these uncertainties. This sort of characterization indeed turns out to be crucial for the relevant analysis of any simulation outputs or the correct interpretation of operational forecasting schemes. In this context, the tools provided by the Bayesian theory play a crucial role since they encompass a variety of methodologies to model and process uncertainty. As a consequence, the Bayesian paradigm has already been present in many contributions of the Fluminance group

in the last years and will remain a cornerstone of the new methodologies investigated by the team in the domain of uncertainty characterization.

This wide theme of research problems is a central topic in our research group. As a matter of fact, such a coupling may rely on adequate instantaneous motion descriptors extracted with the help of the techniques studied in the first research axis of the FLUMINANCE group. In the same time, this coupling is also essential with respect to visual flow control studies explored in the third theme. The coupling between a dynamics and data, designated in the literature as a Data Assimilation issue, can be either conducted with optimal control techniques [44], [45] or through stochastic filtering approaches [37], [40]. These two frameworks have their own advantages and deficiencies. We rely indifferently on both approaches.

3.3. Optimization and control of fluid flows with visual servoing

Fluid flow control is a recent and active research domain. A significant part of the work carried out so far in that field has been dedicated to the control of the transition from laminarity to turbulence. Delaying, accelerating or modifying this transition is of great economical interest for industrial applications. For instance, it has been shown that for an aircraft, a drag reduction can be obtained while enhancing the lift, leading consequently to limit fuel consumption. In contrast, in other application domains such as industrial chemistry, turbulence phenomena are encouraged to improve heat exchange, increase the mixing of chemical components and enhance chemical reactions. Similarly, in military and civilians applications where combustion is involved, the control of mixing by means of turbulence handling rouses a great interest, for example to limit infra-red signatures of fighter aircraft.

Flow control can be achieved in two different ways: passive or active control. Passive control provides a permanent action on a system. Most often it consists in optimizing shapes or in choosing suitable surfacing (see for example [33] where longitudinal riblets are used to reduce the drag caused by turbulence). The main problem with such an approach is that the control is, of course, inoperative when the system changes. Conversely, in active control the action is time varying and adapted to the current system's state. This approach requires an external energy to act on the system through actuators enabling a forcing on the flow through for instance blowing and suction actions [52], [39]. A closed-loop problem can be formulated as an optimal control issue where a control law minimizing an objective cost function (minimization of the drag, minimization of the actuators power, etc.) must be applied to the actuators [30]. Most of the works of the literature indeed comes back to open-loop control approaches [47], [41], [46] or to forcing approaches [38] with control laws acting without any feedback information on the flow actual state. In order for these methods to be operative, the model used to derive the control law must describe as accurately as possible the flow and all the eventual perturbations of the surrounding environment, which is very unlikely in real situations. In addition, as such approaches rely on a perfect model, a high computational costs is usually required. This inescapable pitfall has motivated a strong interest on model reduction. Their key advantage being that they can be specified empirically from the data and represent quite accurately, with only few modes, complex flows' dynamics. This motivates an important research axis in the Fluminance group.

Another important part of the works conducted in Fluminance concerns the study of closed-loop approaches, for which the convergence of the system to a target state is ensured even in the presence of errors (related either to the flow model, the actuators, or the sensors) [35]. However, designing a closed loop control law requires the use of sensors that are both non-intrusive, accurate and adapted to the time and spacial scales of the phenomenon to monitor. Such sensors are unfortunately hardly available in the context of flow control. The only sensors currently used are wall sensors located in a limited set of measurement points [31], [34]. The difficulty is then to reconstruct the entire state of the controlled system from a model based only on the few measurements available on the walls [43]. Instead of relying on sparse measurements, we propose to use denser features estimated from images. With the capabilities of up-to-date imaging sensors, we can expect an improved reconstruction of the flow (both in space and time) enabling the design of efficient image based control laws. This formulation is referred to as visual servoing control scheme.

Visual servoing is a widely used technique for robot control. It consists in using data provided by a vision sensor for controlling the motions of a robot [32]. This technique, historically embedded in the larger domain of sensor-based control [48], can be properly used to control complex robotic systems or, as we showed it recently, flows [51].

Classically, to achieve a visual servoing task, a set of visual features, $\mathbf{s}$, has to be selected from visual measurements, $\mathbf{m}$, extracted from a current image. A control law is then designed so that these visual features reach a desired value, $\mathbf{s}^*$, related to the target state of the system. The control principle consists in regulating to zero the error vector: $\mathbf{e} = \mathbf{s} - \mathbf{s}^*$. To build the control law, the knowledge of the so-called *interaction matrix* $\mathbf{L_s}$ is usually required. This matrix links the time variation of $\mathbf{s}$ to the signal command $\mathbf{u}$. However, computing this matrix in the context of flow control is far more complex than in the case of robot control as flows are associated to chaotic nonlinear systems living in infinite dimensional spaces. As such, it is possible to formalize the model through a Galerkin projection in terms of an ODE system for which classical control laws can be applied. It is also possible to express the system with finite difference approximations and to use discrete time control algorithms amenable to modern micro-controllers. Alternatively, one may develop control methods directly on the infinite dimensional system and then finally discretize the resulting process for implementation purpose. Each approach has its own advantages and drawbacks. For the first two, known control methods can be used at the expense of a great sensibility to space discretization. The last one is less sensitive to discretization errors but more difficult to set up. These practical issues and their related theoretical difficulties make this study a very interesting field of research.

GALEN Project-Team

3. Research Program

3.1. Shape, Grouping and Recognition

A general framework for the fundamental problems of image segmentation, object recognition and scene analysis is the interpretation of an image in terms of a set of symbols and relations among them. Abstractly stated, image interpretation amounts to mapping an observed image, X to a set of symbols Y. Of particular interest are the symbols Y^* that *optimally explain the underlying image*, as measured by a scoring function s that aims at distinguishing correct (consistent with human labellings) from incorrect interpretations:

$$Y^* = \text{argmax}_Y s(X, Y) \tag{70}$$

Applying this framework requires (a) identifying which symbols and relations to use (b) learning a scoring function s from training data and (c) optimizing over Y in Eq. 1 .

Applying this framework requires (a) identifying which symbols and relations to use for image and object representation (b) learning a scoring function s from training data and (c) optimizing over Y in Eq. 1 . One of the main themes of our work is the development of methods that jointly address (a,b,c) in a shape-grouping framework in order to reliably extract, describe, model and detect shape information from natural and medical images. A principal motivation for using a shape-based framework is the understanding that shape- and more generally, grouping- based representations can go all the way from image features to objects. Regarding aspect (a), image representation, we cater for the extraction of image features that respect the shape properties of image structures. Such features are typically constructed to be purely geometric (e.g. boundaries, symmetry axes, image segments), or appearance-based, such as image descriptors. The use of machine learning has been shown to facilitate the robust and efficient extraction of such features, while the grouping of local evidence is known to be necessary to disambiguate the potentially noisy local measurements. In our research we have worked on improving feature extraction, proposing novel blends of invariant geometric- and appearance- based features, as well as grouping algorithms that allow for the efficient construction of optimal assemblies of local features.

Regarding aspect (b) we have worked on learning scoring functions for detection with deformable models that can exploit the developed low-level representations, while also being amenable to efficient optimization. Our works in this direction build on the graph-based framework to construct models that reflect the shape properties of the structure being modeled. We have used discriminative learning to exploit boundary- and symmetry-based representations for the construction of hierarchical models for shape detection, while for medical images we have developed methods for the end-to-end discriminative training of deformable contour models that combine low-level descriptors with contour-based organ boundary representations.

Regarding aspect (c) we have developed algorithms which implement top-down/bottom-up computation both in deterministic and stochastic optimization. The main idea is that 'bottom-up', image-based guidance is necessary for efficient detection, while 'top-down', object-based knowledge can disambiguate and help reliably interpret a given image; a combination of both modes of operation is necessary to combine accuracy with efficiency. In particular we have developed novel techniques for object detection that employ combinatorial optimization tools (A* and Branch-and-Bound) to tame the combinatorial complexity, achieving a best-case performance that is logarithmic in the number of pixels.

In the long run we aim at scaling up shape-based methods to 3D detection and pose estimation and large-scale object detection. One aspect which seems central to this is the development of appropriate mid-level representations. This is a problem that has received increased interest lately in the 2D case and is relatively mature, but in 3D it has been pursued primarily through ad-hoc schemes. We anticipate that questions pertaining to part sharing in 3D will be addressed most successfully by relying on explicit 3D representations. On the one hand depth sensors, such as Microsoft's Kinect, are now cheap enough to bring surface modeling and matching into the mainstream of computer vision - so these advances may be directly exploitable at test time for detection. On the other hand, even if we do not use depth information at test time, having 3D information can simplify the modeling task during training. In on-going work with collaborators we have started exploring combinations of such aspects, namely (i) the use of surface analysis tools to match surfaces from depth sensors (ii) using branch-and-bound for efficient inference in 3D space and (iii) groupwise-registration to build statistical 3D surface models. In the coming years we intend to pursue a tighter integration of these different directions for scalable 3D object recognition.

3.2. Machine Learning & Structured Prediction

The foundation of statistical inference is to learn a function that minimizes the expected loss of a prediction with respect to some unknown distribution

$$\mathcal{R}(f) = \int \ell(f, x, y) dP(x, y), \tag{71}$$

where $\ell(f, x, y)$ is a problem specific loss function that encodes a penalty for predicting $f(x)$ when the correct prediction is y. In our case, we consider x to be a medical image, and y to be some prediction, e.g. the segmentation of a tumor, or a kinematic model of the skeleton. The loss function, ℓ, is informed by the costs associated with making a specific misprediction. As a concrete example, if the true spatial extent of a tumor is encoded in y, $f(x)$ may make mistakes in classifying healthy tissue as a tumor, and mistakes in classifying diseased tissue as healthy. The loss function should encode the potential physiological damage resulting from erroneously targeting healthy tissue for irradiation, as well as the risk from missing a portion of the tumor.

A key problem is that the distribution P is unknown, and any algorithm that is to estimate f from labeled training examples must additionally make an implicit estimate of P. A central technology of empirical inference is to approximate $\mathcal{R}(f)$ with the empirical risk,

$$\mathcal{R}(f) \approx \widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, x_i, y_i), \tag{72}$$

which makes an implicit assumption that the training samples (x_i, y_i) are drawn i.i.d. from P. Direct minimization of $\widehat{\mathcal{R}}(f)$ leads to overfitting when the function class $f \in \mathcal{F}$ is too rich, and regularization is required:

$$\min_{f \in \mathcal{F}} \lambda \Omega(\|f\|) + \widehat{\mathcal{R}}(f), \tag{73}$$

where Ω is a monotonically increasing function that penalizes complex functions.

Equation (4) is very well studied in classical statistics for the case that the output, $y \in \mathcal{Y}$, is a binary or scalar prediction, but this is not the case in most medical imaging prediction tasks of interest. Instead, complex interdependencies in the output space leads to difficulties in modeling inference as a binary prediction problem. One may attempt to model e.g. tumor segmentation as a series of binary predictions at each voxel in a medical image, but this violates the i.i.d. sampling assumption implicit in Equation (3). Furthermore, we typically gain performance by appropriately modeling the inter-relationships between voxel predictions, e.g. by incorporating pairwise and higher order potentials that encode prior knowledge about the problem domain. It is in this context that we develop statistical methods appropriate to structured prediction in the medical imaging setting.

3.3. Self-Paced Learning with Missing Information

Many tasks in artificial intelligence are solved by building a model whose parameters encode the prior domain knowledge and the likelihood of the observed data. In order to use such models in practice, we need to estimate its parameters automatically using training data. The most prevalent paradigm of parameter estimation is supervised learning, which requires the collection of the inputs x_i and the desired outputs y_i. However, such an approach has two main disadvantages. First, obtaining the ground-truth annotation of high-level applications, such as a tight bounding box around all the objects present in an image, is often expensive. This prohibits the use of a large training dataset, which is essential for learning the existing complex models. Second, in many applications, particularly in the field of medical image analysis, obtaining the ground-truth annotation may not be feasible. For example, even the experts may disagree on the correct segmentation of a microscopical image due to the similarities between the appearance of the foreground and background.

In order to address the deficiencies of supervised learning, researchers have started to focus on the problem of parameter estimation with data that contains hidden variables. The hidden variables model the missing information in the annotations. Obtaining such data is practically more feasible: image-level labels ('contains car','does not contain person') instead of tight bounding boxes; partial segmentation of medical images. Formally, the parameters $\mathbf{w}$ of the model are learned by minimizing the following objective:

$$\min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}) + \sum_{i=1}^{n} \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \tag{74}$$

Here, $\mathcal{W}$ represents the space of all parameters, n is the number of training samples, $R(\cdot)$ is a regularization function, and $\Delta(\cdot)$ is a measure of the difference between the ground-truth output y_i and the predicted output and hidden variable pair $(y_i(\mathbf{w}), h_i(\mathbf{w}))$.

Previous attempts at minimizing the above objective function treat all the training samples equally. This is in stark contrast to how a child learns: first focus on easy samples ('learn to add two natural numbers') before moving on to more complex samples ('learn to add two complex numbers'). In our work, we capture this intuition using a novel, iterative algorithm called self-paced learning (SPL). At an iteration t, SPL minimizes the following objective function:

$$\min_{\mathbf{w}\in\mathcal{W},\mathbf{v}\in\{0,1\}^n} R(\mathbf{w}) + \sum_{i=1}^{n} v_i \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})) - \mu_t \sum_{i=1}^{n} v_i. \tag{75}$$

Here, samples with $v_i = 0$ are discarded during the iteration t, since the corresponding loss is multiplied by 0. The term μ_t is a threshold that governs how many samples are discarded. It is annealed at each iteration, allowing the learner to estimate the parameters using more and more samples, until all samples are used. Our results already demonstrate that SPL estimates accurate parameters for various applications such as image classification, discriminative motif finding, handwritten digit recognition and semantic segmentation. We will investigate the use of SPL to estimate the parameters of the models of medical imaging applications, such as segmentation and registration, that are being developed in the GALEN team. The ability to handle missing information is extremely important in this domain due to the similarities between foreground and background appearances (which results in ambiguities in annotations). We will also develop methods that are capable of minimizing more general loss functions that depend on the (unknown) value of the hidden variables, that is,

$$\min_{\mathbf{w}\in\mathcal{W},\theta\in\Theta} R(\mathbf{w}) + \sum_{i=1}^{n} \sum_{h_i\in\mathcal{H}} \Pr(h_i|x_i, y_i; \theta) \Delta(y_i, h_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \tag{76}$$

Here, θ is the parameter vector of the distribution of the hidden variables h_i given the input x_i and output y_i, and needs to be estimated together with the model parameters $\mathbf{w}$. The use of a more general loss function will allow us to better exploit the freely available data with missing information. For example, consider the case where y_i is a binary indicator for the presence of a type of cell in a microscopical image, and h_i is a tight bounding box around the cell. While the loss function $\Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w}))$ can be used to learn to classify an image as containing a particular cell or not, the more general loss function $\Delta(y_i, h_i, y_i(\mathbf{w}), h_i(\mathbf{w}))$ can be used to learn to detect the cell as well (since h_i models its location)

3.4. Discrete Biomedical Image Perception

A wide variety of tasks in medical image analysis can be formulated as discrete labeling problems. In very simple terms, a discrete optimization problem can be stated as follows: we are given a discrete set of variables $\mathcal{V}$, all of which are vertices in a graph $\mathcal{G}$. The edges of this graph (denoted by $\mathcal{E}$) encode the variables' relationships. We are also given as input a discrete set of labels $\mathcal{L}$. We must then assign one label from $\mathcal{L}$ to each variable in $\mathcal{V}$. However, each time we choose to assign a label, say, x_{p_1} to a variable p_1, we are forced to pay a price according to the so-called *singleton* potential function $g_p(x_p)$, while each time we choose to assign a pair of labels, say, x_{p_1} and x_{p_2} to two interrelated variables p_1 and p_2 (two nodes that are connected by an edge in the graph $\mathcal{G}$), we are also forced to pay another price, which is now determined by the so called *pairwise* potential function $f_{p_1p_2}(x_{p_1}, x_{p_2})$. Both the singleton and pairwise potential functions are problem specific and are thus assumed to be provided as input.

Our goal is then to choose a labeling which will allow us to pay the smallest total price. In other words, based on what we have mentioned above, we want to choose a labeling that minimizes the sum of all the MRF potentials, or equivalently the MRF energy. This amounts to solving the following optimization problem:

$$\arg\min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}). \tag{77}$$

The use of such a model can describe a number of challenging problems in medical image analysis. However these simplistic models can only account for simple interactions between variables, a rather constrained scenario for high-level medical imaging perception tasks. One can augment the expression power of this model through higher order interactions between variables, or a number of cliques $\{C_i, i \in [1, n] = \{\{p_{i^1}, \cdots, p_{i|C_i|}\}\}$ of order $|C_i|$ that will augment the definition of $\mathcal{V}$ and will introduce hyper-vertices:

$$\arg\min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}) + \sum_{C_i \in \mathcal{E}} f_{p_1 \cdots p_n}(x_{p_{i^1}}, \cdots, p_{x_{i|C_i|}}). \tag{78}$$

where $f_{p_1 \cdots p_n}$ is the price to pay for associating the labels $(x_{p_{i^1}}, \cdots, p_{x_{i|C_i|}})$ to the nodes $(p_1 \cdots p_{i|C_i|})$. Parameter inference, addressed by minimizing the problem above, is the most critical aspect in computational medicine and efficient optimization algorithms are to be evaluated both in terms of computational complexity as well as of inference performance. State of the art methods include deterministic and non-deterministic annealing, genetic algorithms, max-flow/min-cut techniques and relaxation. These methods offer certain strengths while exhibiting certain limitations, mostly related to the amount of interactions which can be tolerated among neighborhood nodes. In the area of medical imaging where domain knowledge is quite strong, one would expect that such interactions should be enforced at the largest scale possible.

GENSCALE Project-Team

3. Research Program

3.1. Introduction

To tackle challenges brought by the processing of huge amount of genomic data, the main strategy of GenScale is to merge the following computer science expertise:

- Data structure;
- Combinatorial optimization;
- Parallelism.

3.2. Data structure

To face the genomic data tsunami, the design of efficient algorithms involves the optimization of memory footprints. A key point is the design of innovative data structures to represent large genomic datasets into computer memories. Today's limitations come from their size, their construction time, or their centralized (sequential) access. Random accesses to large data structures poorly exploit the sophisticated processor cache memory system. New data structures including compression techniques, probabilistic filters, approximate string matching, or techniques to improve spatial/temporal memory access are developed [3].

3.3. Combinatorial optimization

For wide genome analysis, Next Generation Sequencing (NGS) data processing or protein structure applications, the main issue concerns the exploration of sets of data by time-consuming algorithms, with the aim of identifying solutions that are optimal in a predefined sense. In this context, speeding up such algorithms requires acting on many directions: (1) optimizing the search with efficient heuristics and advanced combinatorial optimization techniques [2], [5] or (2) targeting biological sub-problems to reduce the search space [7], [9]. Designing algorithms with adapted heuristics, and able to scale from protein (a few hundreds of amino acids) to full genome (millions to billions of nucleotides) is one of the competitive challenges addressed in the GenScale project.

3.4. Parallelism

The traditional parallelization approach, which consists in moving from a sequential to a parallel code, must be transformed into a direct design and implementation of high performance parallel software. All levels of parallelism (vector instructions, multi-cores, many-cores, clusters, grid, clouds) need to be exploited in order to extract the maximum computing power from current hardware resources [6], [8], [1]. An important specificity of GenScale is to systematically adopt a design approach where all levels of parallelism are potentially considered.

IBIS Project-Team

3. Research Program

3.1. Analysis of qualitative dynamics of gene regulatory networks

Participants: Hidde de Jong [Correspondent], Michel Page.

The dynamics of gene regulatory networks can be modeled by means of ordinary differential equations (ODEs), describing the rate of synthesis and degradation of the gene products as well as regulatory interactions between gene products and metabolites. In practice, such models are not easy to construct though, as the parameters are often only constrained to within a range spanning several orders of magnitude for most systems of biological interest. Moreover, the models usually consist of a large number of variables, are strongly nonlinear, and include different time-scales, which makes them difficult to handle both mathematically and computationally. This has motivated the interest in qualitative models which, from incomplete knowledge of the system, are able to provide a coarse-grained picture of its dynamics.

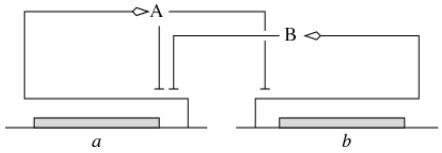
A variety of qualitative modeling formalisms have been introduced over the past decades. Boolean or logical models, which describe gene regulatory and signalling networks as discrete-time finite-state transition systems, are probably most widely used. The dynamics of these systems are governed by logical functions representing the regulatory interactions between the genes and other components of the system. IBIS has focused on a related, hybrid formalism that embeds the logical functions describing regulatory interactions into an ODE formalism, giving rise to so-called piecewise-linear differential equations (PLDEs, Figure 2). The use of logical functions allows the qualitative dynamics of the PLDE models to be analyzed, even in high-dimensional systems. In particular, the qualitative dynamics can be represented by means of a so-called state transition graph, where the states correspond to (hyperrectangular) regions in the state space and transitions between states arise from solutions entering one region from another.

First proposed by Leon Glass and Stuart Kauffman in the early seventies, the mathematical analysis of PLDE models has been the subject of active research for more than four decades. IBIS has made contributions on the mathematical level, in collaboration with the BIOCORE and BIPOP project-teams, notably for solving problems induced by discontinuities in the dynamics of the system at the boundaries between regions, where the logical functions may abruptly switch from one discrete value to another, corresponding to the (in)activation of a gene. In addition, many efforts have gone into the development of the computer tool GENETIC NETWORK ANALYZER (GNA) and its applications to the analysis of the qualitative dynamics of a variety of regulatory networks in microorganisms. Some of the methodological work underlying GNA, notably the development of analysis tools based on temporal logics and model checking, which was carried out with the Inria project-teams CONVEX (ex-VASY) and POP-ART, has implications beyond PLDE models as they apply to logical and other qualitative models as well.

3.2. Inference of gene regulatory networks from time-series data

Participants: Eugenio Cinquemani [Correspondent], Johannes Geiselmann, Hidde de Jong, Julien Demol, Stéphan Lacour, Michel Page, Corinne Pinel, Delphine Ropers, Alberto Soria-Lopéz, Diana Stefan, Valentin Zulkower.

Measurements of the transcriptome of a bacterial cell by means of DNA microarrays, RNA sequencing, and other technologies have yielded huge amounts of data on the state of the transcriptional program in different growth conditions and genetic backgrounds, across different time-points in an experiment. The information on the time-varying state of the cell thus obtained has fueled the development of methods for inferring regulatory interactions between genes. In essence, these methods try to explain the observed variation in the activity of one gene in terms of the variation in activity of other genes. A large number of inference methods have been proposed in the literature and have been successful in a variety of applications, although a number of difficult problems remain.



(a)

$$\dot{x}_a = \kappa_a \, s^-(x_a, \theta_a^2) \, s^-(x_b, \theta_b) - \gamma_a \, x_a$$

$$\dot{x}_b = \kappa_b \, s^-(x_a, \theta_a^1) - \gamma_b \, x_b$$

$$s^+(x, \theta) = \begin{cases} 1, & \text{if } x > \theta \\ 0, & \text{if } x < \theta \end{cases}$$

$$s^-(x, \theta) = 1 - s^+(x, \theta)$$

(b)

Figure 2. (Left) Example of a gene regulatory network of two genes (a and b), each coding for a regulatory protein (A and B). Protein B inhibits the expression of gene a, while protein A inhibits the expression of gene b and its own gene. (Right) PLDE model corresponding to the network in (a). Protein A is synthesized at a rate κ_a, if and only if the concentration of protein A is below its threshold θ_a^2 ($x_a < \theta_a^2$) and the concentration of protein B below its threshold θ_b ($x_b < \theta_b$). The degradation of protein A occurs at a rate proportional to the concentration of the protein itself ($\gamma_a \, x_a$).

Current reporter gene technologies, based on Green Fluorescent Proteins (GFPs) and other fluorescent and luminescent reporter proteins, provide an excellent means to measure the activity of a gene *in vivo* and in real time (Figure 3). The underlying principle of the technology is to fuse the promoter region and possibly (part of) the coding region of a gene of interest to a reporter gene. The expression of the reporter gene generates a visible signal (fluorescence or luminescence) that is easy to capture and reflects the expression of a gene of interest. The interest of the reporter systems is further enhanced when they are applied in mutant strains or combined with expression vectors that allow the controlled induction of any particular gene, or the degradation of its product, at a precise moment during the time-course of the experiment. This makes it possible to perturb the network dynamics in a variety of ways, thus obtaining precious information for network inference.

The specific niche of IBIS in the field of network inference has been the development and application of genome engineering techniques for constructing the reporter and perturbation systems described above, as well as the use of reporter gene data for the reconstruction of gene regulation functions. We have developed an experimental pipeline that resolves most technical difficulties in the generation of reproducible time-series measurements on the population level. The pipeline comes with data analysis software that converts the primary data into measurements of time-varying promoter activities (Section 4.2). In addition, for measuring gene expression on the single-cell level by means of microfluidics and time-lapse fluorescence microscopy, we have established collaborations with groups in Grenoble and Paris. The data thus obtained can be exploited for the structural and parametric identification of gene regulatory networks, for which methods with a solid mathematical foundation are developed, in collaboration with colleagues at ETH Zürich (Switzerland) and the University of Pavia (Italy). The vertical integration of the network inference process, from the construction of the biological material to the data analysis and inference methods, has the advantage that it allows the experimental design to be precisely tuned to the identification requirements.

3.3. Analysis of integrated metabolic and gene regulatory networks

Participants: Eugenio Cinquemani, Hidde de Jong, Johannes Geiselmann, Stéphan Lacour, Yves Markowicz, Manon Morin, Michel Page, Corinne Pinel, Stéphane Pinhal, Delphine Ropers [Correspondent], Diana Stefan, Valentin Zulkower.

The response of bacteria to changes in their environment involves responses on several different levels, from the redistribution of metabolic fluxes and the adjustment of metabolic pools to changes in gene expression. In order to fully understand the mechanisms driving the adaptive response of bacteria, as mentioned above, we need to analyze the interactions between metabolism and gene expression. While often studied in isolation, gene regulatory networks and metabolic networks are closely intertwined. Genes code for enzymes which

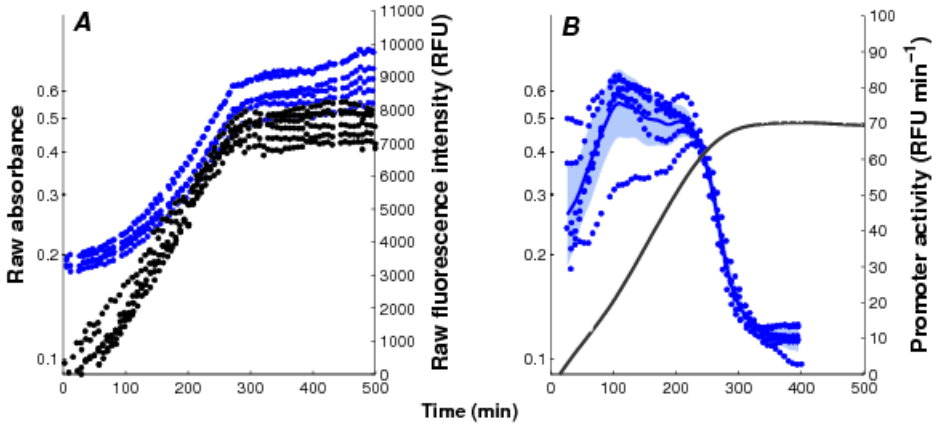


Figure 3. Monitoring of bacterial gene expression in vivo using fluorescent reporter genes (Stefan et al., PLoS Computational Biology, 11(1):e1004028, 2015). The plots show the primary data obtained in a kinetic experiment with E. coli cells, focusing on the expression of the motility gene tar in a mutant background. A: Absorbance (•, black) and fluorescence (•, blue) data, corrected for background intensities, obtained with the ΔcpxR strain transformed with the ptar-gfp reporter plasmid and grown in M9 with glucose. B: Activity of the tar promoter, computed from the primary data. The solid black line corresponds to the mean of 6 replicate absorbance measurements and the shaded blue region to the mean of the promoter activities ± twice the standard error of the mean.

control metabolic fluxes, while the accumulation or depletion of metabolites may affect the activity of transcription factors and thus the expression of enzyme-encoding genes.

The fundamental principles underlying the interactions between gene expressions and metabolism are far from being understood today. From a biological point of view, the problem is quite challenging, as metabolism and gene expression are dynamic processes evolving on different time-scales and governed by different types of kinetics. Moreover, gene expression and metabolism are measured by different experimental methods generating heterogeneous, and often noisy and incomplete data sets. From a modeling point of view, difficult methodological problems concerned with the reduction and calibration of complex nonlinear models need to be addressed.

Most of the work carried out within the IBIS project-team specifically addressed the analysis of integrated metabolic and gene regulatory networks in the context of *E. coli* carbon metabolism (Figure 4). While an enormous amount of data has accumulated on this model system, the complexity of the regulatory mechanisms and the difficulty to precisely control experimental conditions during growth transitions leave many essential questions open, such as the physiological role and the relative importance of mechanisms on different levels of regulation (transcription factors, metabolic effectors, global physiological parameters, ...). We are interested in the elaboration of novel biological concepts and accompanying mathematical methods to grasp the nature of the interactions between metabolism and gene expression, and thus better understand the overall functioning of the system. Moreover, we have worked on the development of methods for solving what is probably the hardest problem when quantifying the interactions between metabolism and gene expression: the estimation of parameters from hetereogeneous and noisy high-throughput data. These problems are tackled in collaboration with experimental groups at Inra/INSA Toulouse and CEA Grenoble, which have complementary experimental competences (proteomics, metabolomics) and biological expertise.

3.4. Natural and engineered control of regulatory networks

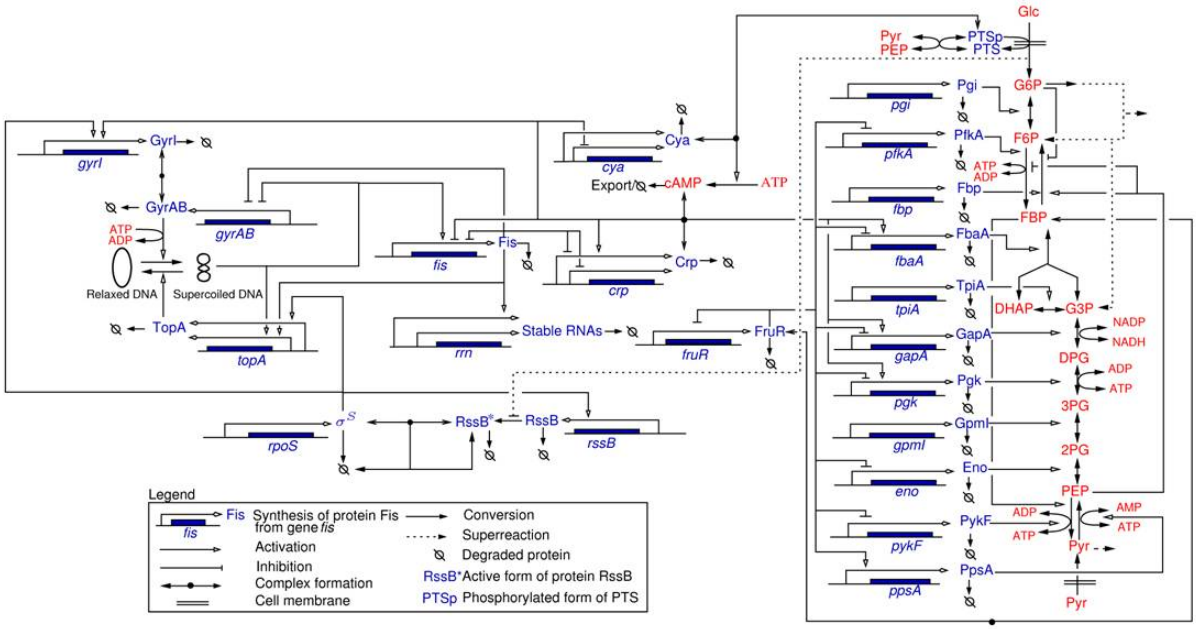


Figure 4. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in E. coli (Baldazzi et al., PLoS Computational Biology, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description of the PTS system, via the phosphorylated and non-phosphorylated form of its enzymes (represented by PTSp and PTS, respectively). The pentose-phosphate pathway (PPP) is not explicitly described but we take into account that a small pool of G6P escapes the upper part of glycolysis. At the level of the global regulators the network includes the control of the DNA supercoiling level, the accumulation of the sigma factor RpoS and the Crp·cAMP complex, and the regulatory role exerted by the fructose repressor FruR.

Participants: Cindy Gomez Balderas-Barillot, Eugenio Cinquemani, Johannes Geiselmann [Correspondent], Edith Grac, Nils Giordano, Hidde de Jong, Stéphan Lacour, Delphine Ropers, Alberto Soria-Lopéz.

In the previously-described objectives, we have focused on identifying complex regulatory networks and gaining a better understanding of how the network dynamics underlies the observable behavior of the cell. Based on the insights thus obtained, a complementary perspective consists in changing the functioning of a bacterial cell towards a user-defined objective, by rewiring and selectively perturbing its regulatory networks. The question how regulatory networks in microorganisms can be externally controlled using engineering approaches has a long history in biotechnology and is receiving much attention in the emerging field of synthetic biology.

Within a number of on-going projects, IBIS is focusing on two different questions. The first concerns the development of growth-rate controllers of bacterial cells. Since the growth rate is the most important physiological parameter in microorganisms, a better understanding of the molecular basis of growth-rate control and the engineering of open-loop and closed-loop growth-rate controllers is of major interest for both fundamental research and biotechnological applications. Second, we are working on the development of methods with a solid foundation in control theory for the real-time control of gene expression. These methods are obviously capital for the above-mentioned design of growth-rate controllers, but they have also been applied in the context of a platform for real-time control of gene expression in cell population and single cells, developed by the Inria project-team CONTRAINTES, in collaboration with a biophysics group at Université Paris Descartes.

KALIFFE Project-Team

3. Research Program

3.1. Numerical schemes for nonlinear kinetic models in an arbitrary geometry

In this part, we want to focus in the numerical approximation of solutions to kinetic equations (microscopic description) set in a complex geometry with different types of boundary conditions. Many numerical schemes have been proposed to approximate the solutions of nonlinear kinetic equations, but few of them are concerned by the treatment of complex geometry and boundary conditions which have a special interest for applications. In this context, classical structured or unstructured meshes already applied in computational fluid dynamics are not appropriate due to the high dimensional property of kinetic problems. In contrast, the Cartesian mesh makes the numerical method efficient and easy to implement. Indeed, in the framework of the Inria-Calvi project, E. Sonnendrücker and his collaborators have developed several families of methods for solving transport equations in a phase space grid with specific applications to plasma physics. These methods are based on the well known semi-Lagrangian methods. The principle is to solve the equation on a phase space grid, for which the grid points are advected with the flow of the transport equation for a time step and interpolated back periodically on the initial grid. The characteristics can be solved either forward or backward in time leading to the forward semi-Lagrangian or backward semi-Lagrangian schemes. These schemes are particularly well suited for uniform Cartesian grid since they are efficient in term of accuracy (high order scheme), stability (not restricted by a CFL condition) and computational cost (fast to locate the transported grid point).

Our aim is now to use all these techniques in the context of complex geometry and for the treatment of boundary conditions. The difficulty is that obviously grid points are usually not located on the physical boundary when using a Cartesian mesh, thus a suitable numerical method to capture the boundary condition on the complex geometry is required. In order to apply numerical methods previously studied, we want to treat separately the transport equation and the boundary conditions in the complex geometry.

Several numerical methods based on Cartesian mesh have been developed in computational fluid dynamics in last decade. Among these methods, the immersed boundary method (IBM), first introduced by Peskin for the study of biological fluid mechanics problems, has attracted considerable attention because of its use of regular Cartesian grid and great simplification of tedious grid generation task. The basic idea of immersed boundary method is that the effect of the immersed boundary on the surrounding fluid is represented through the introduction of forcing terms in the momentum equations. In conservation laws, two major classes immersed boundary like methods can be distinguished on different discretization types. The first class is Cartesian cut-cell method, which is based on a finite volume method. This conceptually simple approach “cuts” solid bodies out of a background Cartesian mesh. Thus we have several polygons (cut-cells) along the boundary. Then the numerical flux at the boundary of these cut-cells are imposed by using the real boundary conditions. This method satisfies well the conservation laws, however to determine the polygons is still a delicate issue.

Here, we will consider another class of method, based on finite difference method. To achieve a high order interior scheme, several ghost points behind the boundary are added. For instance for solving hyperbolic conservations laws, an inverse Lax-Wendroff type procedure is used to impose some artificial values on the ghost points. The interest of this approach is that it preserves all the flexibility of semi-Lagrangian schemes, that is, high order accuracy, resolution in a uniform Cartesian grid and stability. The members of the project involved in this thematic pole are already studying kinetic and related models and will develop this type of numerical schemes focusing on the following goals:

- **Accuracy.** Achieving arbitrary high accuracy for problems with smooth solutions has been a topic of the utmost importance in the recent years and actively studied by a many researchers and groups worldwide. The project team has been investigating such methods for several years and for various PDE models, for steady and unsteady physical problems, using different formulations among which spectral and semi-Lagrangian methods, discontinuous Galerkin methods and finite volume methods.

In all the cases, we consider numerical methods relying on discretization techniques that best fit to the geometrical characteristics of the problems at hand.

- **Robustness.** On the other hand, these methods should also be capable to accurately describe the underlying physical phenomena that may involve highly variable space and time scales. With reference to this characteristic, several strategies are possible: adaptive local refinement/coarsening of the mesh (*i.e* h-adaptivity) and adaptive local variation of the interpolation order (*i.e.* p-adaptivity). Ideally, these two strategies are combined leading to the so-called hp-adaptive methods and that will actually represent an ultimate objective of our research activities. Note that both strategies are all local in nature.
- **Efficiency.** Despite the ever increasing performances of microprocessors, the numerical simulation of realistic $4D$ or $5D$ kinetic problems is hardly performed on a high-end workstation and parallel computing is a mandatory path. Hence, numerical methods must be adapted to the characteristics of modern parallel computing platforms taking into account their heterogeneity and their hierarchical nature (*e.g.* multiple processors and multiple core systems with complex cache and memory hierarchies, possibly augmented with accelerator cards). Appropriate parallelization strategies need to be designed that combine distributed memory and shared memory paradigms *i.e.* MIMD (multiple instruction, multiple data) and SIMD (single instruction, multiple data) programming models.

3.2. Asymptotic Preserving schemes

We develop robust numerical schemes for kinetic equations that also work in the fluid regime. The goal of this part of the project is to propose a new general and systematic strategy that avoids the inversion of the involved time implicit schemes and that allows to apply the microscopic description without any stability constraint on the numerical parameter h.

Development of numerical schemes for stiff problems.

The idea is to combine micro/macro decomposition with penalization techniques for collision operators, leading to completely explicit schemes which are stable in the desired macroscopic limits. The expected schemes should be consistent with the model at both macroscopic and microscopic levels.

However for plasma applications, the Landau-Fokker-Planck operator has a diffusion structure in the velocity variable which induces special behaviors at both mathematical and numerical levels. We will show that the previous methodology can be adapted to overcome the velocity diffusion stiffness in this case. In other words, the obtained numerical schemes are expected to be free of usual diffusion CFL conditions, and will be stable and consistent within the macroscopic limit under consideration. Finally, to decrease the computational cost of the so constructed schemes, which is due to the non local character of the involved collision operator (Boltzmann, Landau, etc), fast computational method for integral operator are needed. On the basis of multigrid and/or Fast Multi-pole Methods, we shall develop appropriate acceleration methods to our context.

Another important task in this project is to extend the above strategy to the context of a self-consistent Poisson or Maxwell equations. Accurate methods based on finite volume schemes will be developed for high field limit. A stiffness raised by the presence of high electromagnetic fields will also be treated in the same spirit. Such problems are also investigated in the IPSO project (M. Lemou, F. Méhats and N. Crouseilles). Here our strategy is based on a suitable operator decomposition coupled with appropriate IMEX schemes.

Stability and accuracy issue. In the framework of Asymptotic-Preserving (AP) schemes, there are few mathematical justifications of stability and uniform accuracy of such approach. A stability analysis has to be performed to rigorously prove that the numerical scheme is stable for small values of small physical parameters even if the time step does not resolve it. This analysis seems to be tricky for fully nonlinear kinetic equations like the Boltzmann equation. Therefore, we focus on simpler models as discrete velocity models (DVM) which have the same properties as the full Boltzmann operator but deal with a finite set of velocity. In this project we are particularly interested by the long time behavior of the numerical solution when it approaches its stationary state. We plan to apply the entropy-entropy dissipation technique to design new numerical approximations. It gives a specific discretization based on finite volume approximation, which allows to control the numerical

entropy production and in some situations, it is often enough to give stability of the numerical solution in the long time asymptotic limit. For general cases, these estimations have to be completed by some discrete functional inequalities.

LEMON Team

3. Research Program

3.1. State of the Art

3.1.1. Shallow Water Models

Shallow Water (SW) wave dynamics and dissipation represent an important research field. This is because shallow water flows are the most common flows in geophysics. In shallow water regions, dispersive effects (non-hydrostatic pressure effects related to strong curvature in the flow streamlines) can become significant and affect wave transformations. The shoaling of the wave (the "steepening" that happens before the breaking) cannot be described with the usual Saint-Venant equations. To model such various evolutions, one has to use more sophisticated models (Boussinesq, Green-Naghdi...). Nowadays, the classical Saint-Venant equations can be solved numerically in an accurate way, allowing the generation of bores and the shoreline motion to be handled, using recent finite-volume or discontinuous-Galerkin schemes. In contrast, very few advanced works regarding the derivation and modern numerical solution of dispersive equations [28], [32], [60] are available in one dimensions, let alone in the multidimensional case. We can refer to [58], [35] for some linear dispersive equations, treated with finite-element methods, or to [32] for the first use of advanced high-order compact finite-volume methods for the Serre equations. Recent work undertaken during the ANR MathOCEAN [28] lead to some new 1D fully nonlinear and weakly dispersive models (Green-Naghdi like models) that allow to accurately handle the nonlinear waves transformations. High order accuracy numerical methods (based on a second-order splitting strategy) have been developed and implemented, raising a new and promising 1D numerical model. However, there is still a lack of new development regarding the multidimensional case.
In shallow water regions, depending on the complex balance between non-linear effects, dispersive effects and energy dissipation due to wave breaking, wave fronts can evolve into a large range of bore types, from purely breaking to purely undular bore. Boussinesq or Green-Naghdi models can handle these phenomena [26] . However, these models neglect the wave overturning and the associated dissipation, and the dispersive terms are not justified in the vicinity of the singularity. Previous numerical studies concerning bore dynamics using depth-averaged models have been devoted to either purely broken bores using NSW models [29], or undular bores using Boussinesq-type models [39]. Let us also mention [37] for tsunami modeling and [36], [48] for the dam-break problem. A model able to reproduce the various bore shapes, as well as the transition from one type of bore to another, is required. A first step has been made with the one-dimensional code [28], [56]. The SWASH project led by Zijlema at Delft [60] addresses the same issues.

3.1.2. Open boundary conditions and coupling algorithms

For every model set in a bounded domain, there is a need to consider boundary conditions. When the boundaries correspond to a modeling choice rather than to a physical reality, the corresponding boundary conditions should not create spurious oscillations or other unphysical behaviour at the artificial boundary. Such conditions are called **open boundary conditions** (OBC). They have been widely studied by applied mathematicians since the pionneering work of [38] on transparent boundary conditions. Deep studies of these operators have been performed in the case of linear equations, [43], [27], [53]. Unfortunately, in the case of geophysical fluid dynamics, this theory leads to nonlocal conditions (even in linear cases) that are not usable in numerical models. Most of current models (including high quality operational ones) modestly use a *no flux* condition (namely an homogeneous Neumann boundary condition) when a free boundary condition is required. But in many cases, Neumann homogeneous conditions are a very poor approximation of the exact transparent conditions. Hence the need to build higher order approximations of these conditions that remain numerically tractable.

Numerous physical processes are involved in coastal modeling, each of them depending on others (surface winds for coastal oceanography, sea currents for sandbars dynamics, etc.). Connecting two (or more) model solutions at their interface is a difficult task, that is often addressed in a simplified way from the mathematical viewpoint: this can be viewed as the one and only iteration of an iterative process. This results with a low quality coupled system, which could be improved either with additional iterations, and/or thanks to the improvement of interface boundary conditions and the use of OBC (see above). Promising results have been obtained in the framework of **ocean-atmosphere coupling** (in a simplified modeling context) in [49], where the use of advanced coupling techniques (based on domain decomposition algorithm) are introduced.

3.1.3. A need for upscaled shallow water models.

The mathematical modeling of **fluid-biology** coupled systems in lagoon ecosystems requires one or several water models. It is of course not necessary (and not numerically feasible) to use accurate non-hydrostatic turbulent models to force the biological processes over very long periods of time. There is a compromise to be reached between accurate (but untractable) fluid models such as the Navier-Stokes equations and simple (but imprecise) models such as [40].
In urbanized coastal zones, upscaling is also a key issue. This stems not only from the multi-scale aspects dealt with in the previous subsection, but also from modeling efficiency considerations.

The typical size of the relevant hydraulic feature in an urban area is between 0.1 m and 1.0 m, while the size of an urban area usually ranges from 10^3 m to 10^4 m. Refined flow computations (e.g. in simulating the impact of a tsunami) over entire coastal conurbations using a 2D horizontal model thus require 10^6 to 10^9 elements. From an engineering perspective, this makes both the CPU and man-supervised mesh design efforts unaffordable in the present state of technology.

Upscaling provides an answer to this problem by allowing macroscopic equations to be derived from the small-scale governing equations. The powerful, multiple scale expansion-based homogeneization technique [25], [24], [52] has been applied successfully to flow and transport upscaling in porous media, but its use is subordinated to the stringent assumptions of (i) the existence of a Representative Elementary Volume (REV), (ii) the scale separation principle, and (iii) the process is not purely hyperbolic at the microscopic scale, otherwise precluding the study of transient solutions [25]. Unfortunately, the REV has been shown recently not to exist in urban areas [42]. Besides, the scale separation principle is violated in the case of sharp transients (such as tsunami waves) impacting urban areas because the typical wavelength is of the same order of magnitude as the microscopic detail (the street/block size). Moreover, 2D shallow water equations are essentially hyperbolic, thus violating the third assumption.

These hurdles are overcome by averaging approaches. Single porosity-based, macroscopic shallow water models have been proposed [34], [41], [44] and applied successfully to urban flood modeling scale experiments [41], [50], [55]. They allow the CPU time to be divided by 10 to 100 compared to classical 2D shallow water models. Recent extensions of these models have been proposed in the form of integral porosity [54] and multiple porosity [42] shallow water models.

3.2. Scientific Objectives

Our main challenge is: build and couple elementary models in coastal areas to improve their capacity to simulate complex dynamics. This challenge consists of three principal scientific objectives. First of all, each of the elementary models has to be consistently developed (regardless of boundary conditions and interactions with other processes). Then open boundary conditions (for the simulation of physical processes in bounded domains) and links between the models (interface conditions) have to be identified and formalized. Finally, models and boundary conditions (*i.e.* coupled systems) should be proposed, analyzed and implemented in a common platform.

3.2.1. Single process models and boundary conditions

The time-evolution of a water flow in a three-dimensional computational domain is classically modeled by Navier-Stokes equations for incompressible fluids. Depending on the physical description of the considered domain, these equations can be simplified or enriched. Consequently, there are **numerous water dynamics models** that are derived from the original Navier-Stokes equations, such as primitive equations, shallow water equations (see [33]), Boussinesq-type dispersive models [26]), etc. The aforementioned models have **very different mathematical natures**: hyperbolic *vs* parabolic, hydrostatic *vs* non-hydrostatic, inviscid *vs* viscous, etc. They all carry nonlinearities that make their mathematical study (existence, uniqueness and regularity of weak and/or strong solutions) highly challenging (not to speak about the $1M Clay competition for the 3D Navier Stokes equations, which may remain open for some time).

The objective is to focus on the mathematical and numerical modeling of models adapted to **nearshore dynamics**, accounting for complicated wave processes. There exists a large range of models, from the shallow water equations (eventually weakly dispersive) to some fully dispersive deeper models. All these models can be obtained from a suitable asymptotic analysis of the water wave equations (Zakharov formulation) and if the theoretical study of these equations has been recently investigated [47], there is still some serious numerical challenges. So we plan to focus on the derivation and implementation of robust and high order discretization methods for suitable two dimensional models, including enhanced fully nonlinear dispersive models and fully dispersive models, like the Matsuno-generalized approach proposed in [46]. Another objective is to study the shallow water dispersive models without any irrotational flow assumption. Such a study would be of great interest for the study of nearshore circulation (wave induced rip currents).

For obvious physical and/or computational reasons, our models are set in bounded domains. Two types of boundaries are considered: physical and mathematical. Physical boundaries are materialized by an existing interface (atmosphere/ocean, ocean/sand, shoreline, etc.) whereas mathematical boundaries appear with the truncation of the domain of interest. In the latter case, **open boundary conditions** are mandatory in order not to create spurious reflexions at the boundaries. Such boundary conditions being nonlocal and impossible to use in practice, we shall look for approximations. We shall obtain them thanks to the asymptotic analysis of the (pseudo-differential) boundary operators with respect to small parameters (viscosity, domain aspect ratio, Rossby number, etc.). Naturally, we **will seek the boundary conditions leading to the best compromise** between mathematical well-posedness and physical consistency. This will make extensive use of the mathematical theory of **absorbing operators** and their approximations [38].

3.2.2. Coupled systems

The Green-Naghdi equations provide a correct description of the waves up to the breaking point while the Saint-Venant equations are more suitable for the description of the surf zone (i.e. after the breaking). Therefore, the challenge here is first to **design a coupling strategy** between these two systems of equations, first in a simplified one-dimensional case, then to the two-dimensional case both on cartesian and unstructured grids. High order accuracy should be achieved through the use of flexible Discontinuous-Galerkin methods.
Additionally, we will couple our weakly dispersive shallow water models to other fully dispersive deeper water models. We plan to mathematically analyze the coupling between these models. In a first step, we have to understand well the mixed problem (initial and boundary conditions) for these systems. In a second step, these new mathematical development have to be embedded within a numerically efficient strong coupling approach. The deep water model should be fully dispersive (solved using spectral methods, for instance) and the shallow-water model will be, in a first approach, the Saint-Venant equations. Then, when the 2D extension of the currently developped Green-Naghdi numerical code will be available, the improved coupling with a weakly dispersive shallow water model should be considered.

In the context of Schwarz relaxation methods, usual techniques can be seen as the first iteration (not converged) of an iterative algorithm. Thanks to the work performed on efficient boundary conditions, we shall **improve the quality of current coupling algorithms**, allowing for qualitatively satisfying solutions **with a reduced computational cost** (small number of iterations).

We are also willing to explore the role of geophysical processes on some biological ones. For example, the design of optimal shellfish farms relies on confinement maps and plankton dynamics, which strongly depend on long-time averaged currents. Equations that model the time evolution of species in a coastal ecosystem are relatively simple from a modeling viewpoint: they mainly consist of ODEs, and possibly advection-diffusion equations. The issue we want to tackle is the choice of the fluid model that should be coupled to them, accounting for the important time scales discrepancy between biological (evolution) processes and coastal fluid dynamics. Discrimination criteria between refined models (such as turbulent Navier-Stokes) and cheap ones (see [40]) will be proposed.

Coastal processes evolve at very different time scales: atmosphere (seconds/minutes), ocean (hours), sediment (months/years) and species evolution (years/decades). Their coupling can be seen as a *slow-fast* dynamical system, and a naïve way to couple them would be to pick the smallest time-step and run the two models together: but the computational cost would then be way too large. Consequently **homogenization techniques or other upscaling methods** should be used in order to account for these various time scales at an affordable computational cost. The research objectives are the following:

- So far, the proposed upscaled models have been validated against theoretical results obtained from refined 2D shallow water models and/or very limited data sets from scale model experiments. The various approaches proposed in the literature [30], [31], [34], [41], [42], [44], [50], [54], [55] have not been compared over the same data sets. Part of the research effort will focus on the extensive validation of the models on the basis of scale model experiments. Active cooperation will be sought with a number of national and international Academic partners involved in urban hydraulics (UCL Louvain-la-Neuve, IMFS Strasbourg, Irvine University California) with operational experimental facilities.
- Upscaling of source terms. Two types of source terms play a key role in shallow water models: geometry-induced source terms (arising from the irregular bathymetry) and friction/turbulence-induced energy loss terms. In all the upscaled shallow water models presented so far, only the large scale effects of topographical variations have been upscaled. In the case of wetting/drying phenomena and small depths (e.g. the *Camargue* tidal flats), however, it is forseen that subgrid-scale topographic variations may play a predominant role. Research on the integration of subgrid-scale topography into macrosocopic shallow water models is thus needed. Upscaling of friction/turbulence-induced head loss terms is also a subject for research, with a number of competing approaches available from the literature [41], [42], [54], [57].
- Upscaling of transport processes. The upscaling of surface pollutant transport processes in the urban environment has not been addressed so far in the literature. Free surface flows in urban areas are characterized by strongly variable (in both time and space) flow fields. Dead/swirling zones have been shown to play a predominant role in the upscaling of the flow equations [42], [54]. Their role is expected to be even stronger in the upscaling of contaminant transport. While numerical experiments indicate that the microscopic hydrodynamic time scales are small compared to the macroscopic time scales, theoretical considerations indicate that this may not be the case with scalar transport. Trapping phenomena at the microscopic scale are well-known to be upscaled in the form of fractional dynamics models in the long time limit [45], [51]. The difficulty in the present research is that upscaling is not sought only for the long time limit but also for all time scales. Fractional dynamics will thus probably not suffice to a proper upscaling of the transport equations at all time scales.

3.2.3. Numerical platform

As a long term objective, the team shall create a common architecture for existing codes, and also the future codes developed by the project members, to offer a simplified management of various evolutions and a single and well documented tool for our partners. It will aim to be self-contained including pre and post-processing tools (efficient meshing approaches, GMT and VTK libraries), but must of course also be opened to user's suggestions, and account for existing tools inside and outside Inria. This numerical platform will be dedicated to the simulation of all the phenomena of interest, including flow propagation, sediment evolution, model

coupling on large scales, from deep water to the shoreline, including swell propagation, shoaling, breaking and run-up. This numerical platform clearly aims at becoming a reference software in the community. It should be used to **develop a specific test case** around Montpellier which embeds many processes and their mutual interactions: from the *Camargue* (where the Rhône river flows into the Mediterranean sea) to the *Étang de Thau* (a wide lagoon where shellfishes are plentiful), **all the processes studied in the project occur in a 100km wide region**, including of course the various hydrodynamics regimes (from the deep sea to the shoaling, surf and swash zones) and crucial morphodynamic issues (*e.g.* in the town of Sete).

LIFEWARE Team

3. Research Program

3.1. Computational Systems Biology

Bridging the gap between the complexity of biological systems and our capacity to model and **quantitatively predict system behaviors** is a central challenge in systems biology. We believe that a deeper understanding of the concept and theory of biochemical computation is necessary to tackle that challenge. Progress in the theory is necessary for scaling, and enabling the application of static analysis, module identification and decomposition, model reductions, parameter search, and model inference methods to large biochemical reaction systems. A measure of success on this route will be the production of better computational modeling tools for elucidating the complex dynamics of natural biological processes, designing synthetic biological circuits and biosensors, developing novel therapy strategies, and optimizing patient-tailored therapeutics.

Progress on the **coupling of models to data** is also necessary. Our approach based on quantitative temporal logics provides a powerful framework for formalizing experimental observations and using them as formal specification in model building. Key to success is a tight integration between *in vivo* and *in silico* work, and on the mixing of dry and wet experiments, enabled by novel biotechnologies. In particular, the use of microfluidic devices makes it possible to measure behaviors at both single-cell and cell population levels *in vivo*, provided innovative modeling, analysis and control methods are deployed *in silico*.

In synthetic biology, while the construction of simple intracellular circuits has shown feasible, the design of larger, **multicellular systems** is a major open issue. In engineered tissues for example, the behavior results from the subtle interplay between intracellular processes (signal transduction, gene expression) and intercellular processes (contact inhibition, gradient of diffusible molecule), and the question is how should cells be genetically modified such that the desired behavior robustly emerges from cell interactions.

3.2. Modeling of Cellular Processes

Since nearly two decades, a significant interest has grown for getting a quantitative understanding of the functioning of biological systems at the cellular level. Given their complexity, proposing a model accounting for the observed cell responses, or better, predicting novel behaviors, is now regarded as an essential step to validate a proposed mechanism in systems biology. Moreover, the constant improvement of stimulation and observation tools creates a strong push for the development of methods that provide predictions that are increasingly precise (single cell precision) and robust (complex stimulation profiles). In addition to the widely-used ordinary differential equation modeling framework, stochastic modeling frameworks, such as chemical master equations, and statistic modeling frameworks, such as ensemble models, are increasingly popular, since they enable to capture biological variability.

In all cases, dedicated mathematical and computational approaches are needed for the analysis of the models and their calibration to experimental data. One can notably mention global optimization tools to search for appropriate parameters within large spaces, moment closure approaches to efficiently approximate stochastic models, and (stochastic approximations of) the expectation maximization algorithm for the identification of mixed-effects models.

3.3. External Control of Cell Processes

External control has been employed since many years to regulate culture growth and other physiological properties. Recently, taking inspiration from developments in synthetic biology, closed loop control has been applied to the regulation of intracellular processes. Such approaches offer unprecedented opportunities to investigate how a cell process dynamical information by maintaining it around specific operating points or driving it out of its standard operating conditions. They can also be used to complement and help the development of synthetic biology through the creation of hybrid systems resulting from the interconnection of in vivo and in silico computing devices.

In collaboration with Pascal Hersen (CNRS MSC lab), we developed a platform for gene expression control that enables to control protein concentrations in yeast cells. This platform integrates microfluidic devices enabling long-term observation and rapid change of the cells environment, microscopy for single cell measurements, and software for real-time signal quantification and model based control. We demonstrated recently that this platform enables controlling the level of a fluorescent protein in cells with unprecedented accuracy and for many cell generations [0].

3.4. Chemical Reaction Network Theory

Feinberg's chemical reaction network theory and Thomas's influence network analyses provide sufficient and/or necessary structural conditions for the existence of multiple steady states and oscillations in regulatory networks, which can be predicted by static analyzers without making any simulation. In this domain, most of our work consists in analyzing the interplay between the **structure** (Petri net properties, influence graph, subgraph epimorphisms) and the **dynamics** (Boolean, CTMC, ODE, time scale separations) of biochemical reaction systems. In particular, our study of influence graphs of reaction systems, our generalization of Thomas' conditions of multi-stationarity and Soulé's proof to reaction systems [0], the inference of reaction systems from ODEs [8], the computation of structural invariants by constraint programming techniques, and the analysis of model reductions by subgraph epimorphisms [9], now provide solid ground for developing static analyzers, using them on a large scale in systems biology, and elucidating modules.

3.5. Logical Paradigm for Systems Biology

Our group was among the first ones in 2002 to apply **model-checking** methods to systems biology in order to reason on large molecular interaction networks, such as Kohn's map of the mammalian cell cycle (800 reactions over 500 molecules) [0]. The logical paradigm for systems biology that we have subsequently developped for quantitative models can be summarized by the following identifications :

$$\textit{biological model} = \textit{transition system},$$

$$\textit{biological property} = \textit{temporal logic formula},$$

$$\textit{model validation} = \textit{model-checking},$$

$$\textit{model inference} = \textit{constraint solving}.$$

In particular, the definition of a continuous satisfaction degree for **first-order temporal logic** formulae with constraints over the reals, was the key to generalize this approach to quantitative models, opening up the field of model-checking to model optimization. This line of research continues with the development of patterns with efficient solvers [20], [19] and their generalization to handle stochastic effects.

3.6. Constraint solving and optimization

Optimization methods are important in our research. On the one hand, static analysis of biochemical reaction networks involves solving hard combinatorial optimization problems, for which **constraint programming** techniques have shown particularly successful, often beating dedicated algorithms and allowing to solve large instances from model repositories. On the other hand, parameter search and model calibration problems involve similarly solving hard continuous optimization problems, for which **evolutionary algorithms** such as the covariance matrix evolution strategy (CMA-ES) [0] has shown to provide best results in our context, for

[0] Jannis Uhlendorf, Agnés Miermont, Thierry Delaveau, Gilles Charvin, François Fages, Samuel Bottani, Grégory Batt, Pascal Hersen. Long-term model predictive control of gene expression at the population and single-cell levels. Proceedings of the National Academy of Sciences USA, 109(35):14271–14276, 2012.

[0] Sylvain Soliman. A stronger necessary condition for the multistationarity of chemical reaction networks. Bulletin of Mathematical Biology, 75(11):2289–2303, 2013.

[0] N. Chabrier-Rivier, M. Chiaverini, V. Danos, F. Fages, V. Schächter. Modeling and querying biochemical interaction networks. Theoretical Computer Science, 325(1):25–44, 2004.

[0] N. Hansen, A. Ostermeier (2001). Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation, 9(2) pp. 159–195.

up to 100 parameters, for building challenging quantitative models, gaining model-based insights, revisiting admitted assumptions and contributing to biological knowledge [0]

[0]Domitille Heitzler, Guillaume Durand, Nathalie Gallay, Aurélien Rizk, Seungkirl Ahn, Jihee Kim, Jonathan D. Violin, Laurence Dupuy, Christophe Gauthier, Vincent Piketty, Pascale Crépieux, Anne Poupon, Frédérique Clément, François Fages, Robert J. Lefkowitz, Eric Reiter. Competing G protein-coupled receptor kinases balance G protein and β-arrestin signaling. Molecular Systems Biology, 8(590), 2012.

M3DISIM Team

3. Research Program

3.1. Multi-scale modeling and coupling mechanisms for biomechanical systems, with mathematical and numerical analysis

Over the past decade, we have laid out the foundations of a multi-scale 3D model of the cardiac mechanical contraction responding to electrical activation. Several collaborations have been crucial in this enterprise, see below references. By integrating this formulation with adapted numerical methods, we are now able to represent the whole organ behavior in interaction with the blood during complete heart beats. This subject was our first achievement to combine a deep understanding of the underlying physics and physiology and our constant concern of proposing well-posed mathematical formulations and adequate numerical discretizations. In fact, we have shown that our model satisfies the essential thermo-mechanical laws, and in particular the energy balance, and proposed compatible numerical schemes that – in consequence – can be rigorously analyzed, see [5]. In the same spirit, we have recently formulated a poromechanical model adapted to the blood perfusion in the heart, hence precisely taking into account the large deformation of the mechanical medium, the fluid inertia and moving domain, and so that the energy balance between fluid and solid is fulfilled from the model construction to its discretization, see [16].

3.2. Inverse problems with actual data – Fundamental formulation, mathematical analysis and applications

A major challenge in the context of biomechanical modeling – and more generally in modeling for life sciences – lies in using the large amount of data available on the system to circumvent the lack of absolute modeling ground truth, since every system considered is in fact patient-specific, with possibly non-standard conditions associated with a disease. We have already developed original strategies for solving this particular type of inverse problems by adopting the observer stand-point. The idea we proposed consists in incorporating to the classical discretization of the mechanical system an estimator filter that can use the data to improve the quality of the global approximation, and concurrently identify some uncertain parameters possibly related to a diseased state of the patient, see [6], [7], [8]. Therefore, our strategy leads to a coupled model-data system solved similarly to a usual PDE-based model, with a computational cost directly comparable to classical Galerkin approximations. We have already worked on the formulation, the mathematical and numerical analysis of the resulting system – see [3] – and the demonstration of the capabilities of this approach in the context of identification of constitutive parameters for a heart model with real data, including medical imaging, see [1].

MAGIQUE-3D Project-Team

3. Research Program

3.1. Inverse Problems

- **Inverse scattering problems.** The determination of the shape of an obstacle immersed in a fluid medium from some measurements of the scattered field in the presence of incident waves is an important problem in many technologies such as sonar, radar, geophysical exploration, medical imaging and nondestructive testing. Because of its nonlinear and ill-posed character, this inverse obstacle problem (IOP) is very difficult to solve, especially from a numerical viewpoint. The success of the reconstruction depends strongly on the quantity and quality of the measurements, especially on the aperture (range of observation angles) and the level of noise in the data. Moreover, in order to solve IOP, the understanding of the theory for the associated direct scattering problem and the mastery of the corresponding solution methods are fundamental. Magique-3d is involved in the mathematical and numerical analysis of a direct elasto-acoustic scattering problem and of an inverse obstacle scattering problem. More specifically, the purpose of this research axis is to propose a solution methodology for the IOP based on a regularized Newton-type method, known to be robust and efficient.
- **Depth Imaging in the context of DIP.** The challenge of seismic imaging is to obtain an accurate representation of the subsurface from the solution of the full wave equation that is the best mathematical model according to the time reversibility of its solution. The Reverse Time Migration, [71], is a technique for Imaging that is widely used in the industry. It is an iterative process based on the solution of a collection of wave equations. The high complexity of the propagation medium requires the use of advanced numerical methods, which allows one to solve several wave equations quickly and accurately. Magique-3D is involved in Depth Imaging by the way of a collaboration with TOTAL, in the framework of the research program DIP which has been jointly defined by researchers of MAGIQUE-3D and engineers of TOTAL. In this context, MAGIQUE-3D develops new algorithms in order to improve the RTM.

3.2. Modeling

The main activities of Magique-3D in modeling are the derivation and the analysis of models that are based on mathematical physics and are suggested by geophysical problems. In particular, Magique-3D considers equations of interest for the oil industry and focuses on the development and the analysis of numerical models which are well-adapted to solve quickly and accurately problems set in very large or unbounded domains as it is generally the case in geophysics.

- **High-Order Time Schemes.** Using the full wave equation for migration implies very high computational burdens, in order to get high resolution images. Indeed, to improve the accuracy of the numerical solution, one must significantly reduce the space step, which is the distance between two points of the mesh representing the computational domain. Another solution consists in using high-order finite element methods, which are very accurate even with coarse meshes. However, to take fully advantage of the high-order space discretization, one has to develop also high-order time schemes. The most popular ones for geophysical applications are the modified equation scheme [75], [83] and the ADER scheme [79]. Both rely on the same principle, which consists in applying a Taylor expansion in time to the solution of the wave equation. Then, the high-order derivatives with respect to the time are replaced by high order space operators, using the wave equation. Finally, auxiliary variables are introduced in order to transform the differential equation involving high-order operators into a system of differential equation with low order operators. The advantage of this technique is that it leads to explicit time schemes, which avoids the solution of huge linear systems.

The counterpart is that the schemes are only conditionally stable, which means that the time step is constrained by a CFL (Courant-Friedrichs-Levy) condition. The CFL number defines an upper bound for the time step in such a way that the smaller the space step is, the higher the numbers of iterations will be. Magique-3D is working on the construction and the analysis of new explicit and implicit time schemes which have either larger CFL numbers or local CFL numbers. By this way, the computational costs can be reduced without hampering the accuracy of the numerical solution.

- **Finite Element Methods for the time-harmonic wave equation.** As an alternative to Time-Domain Seismic Imaging, geophysicists are more and more interested by Time-Harmonic Seismic Imaging. The drawback of Time Domain Seismic Imaging is that it requires either to store the solution at each time step of the computation, or to perform many solutions to the wave equation. The advantage of Time Harmonic problems is that the solutions can be computed independently for each frequency and the images are produced with only two computations of the wave equation and without storing the solution. The counterpart is that one has to solve a huge linear system, which can not be achieved today when considering realistic 3D elastic media, even with the tremendous progress of Scientific Computing. Discontinuous Galerkin Methods (DGM), which are well-suited for hp-adaptivity, allow the use of coarser meshes without hampering the accuracy of the solution. We are confident that these methods will help us to reduce the size of the linear system to be solved, but they still have to be improved in order to tackle realistic 3D problems. However, there exists many different DGMs, and the choice of the most appropriate one for geophysical applications is still not obvious. Our objectives are **a)** to propose a benchmark in order to test the performances of DGMs for seismic applications and **b)** to improve the most efficient DGMs in order to be able to tackle realistic applications. To these aims, we propose to work in the following directions :

  1. To implement a 2D and 3D solver for time harmonic acoustic and elastodynamic wave equation, based on the Interior Penalty Discontinuous Galerkin Method(IPDGM). The implementation of this solver has started few years ago (see Section 5.1) for solving Inverse Scattering Problems and the results we obtained in 2D let us presage that IPDGM will be well-adapted for geophysical problems.
  2. To develop a new hybridizable DG (HDG) [73] for 2D and 3D elastodynamic equation. Instead of solving a linear system involving the degrees of freedom of all volumetric cells of the mesh, the principle of HDG consists in introducing a Lagrange multiplier representing the trace of the numerical solution on each face of the mesh. Hence, it reduces the number of unknowns of the global linear system and the volumetric solution is recovered thanks to a local computation on each element.
  3. To develop upscaling methods for very heterogeneous media. When the heterogeneities are too small compared with the wavelengths of the waves, it is necessary to use such techniques, which are able to reproduce fine scale effects with computations on coarse meshes only.

  We also intend to consider finite elements methods where the basis functions are not polynomials, but solutions to the time-harmonic wave equations. We have already developed a numerical method based on plane wave basis functions [78]. The numerical results we have obtained on academic test cases showed that the proposed method is not only more stable than the DGM, but also exhibits a better level of accuracy. These results were obtained by choosing the same plane waves for the basis functions of every element of the mesh. We are now considering a new methodology allowing for the optimization of the angle of incidence of the plane waves at the element level.

  Last, we are developing an original numerical methods where the basis functions are fundamental solutions to the Helmholtz equation, such as Bessel or Hankel functions. Moreover, each basis function is not defined element by element but on the whole domain. This allows for reducing the volumetric variational formulation to a surfacic variational formulation.

- **Boundary conditions.** The construction of efficient absorbing boundary conditions (ABC) is very important for solving wave equations. Indeed, wave problems are generally set in unbounded or

very large domains and simulation requires to limit the computational domain by introducing an external boundary, the so-called absorbing boundary. This topic has been a very active research topic during the past twenty years and despite that, efficient ABCs are have still to be designed. Classical conditions are constructed to absorb propagating waves and Magique-3D is investigating the way of improving existing ABCs by introducing the modelling of evanescent and glancing waves. For that purpose, we consider the micro-local derivation of the Dirichlet-to-Neumann operator. The interest of our approach is that the derivation does not depend on the geometry of the absorbing surface.

ABCs have been given up when Perfectly Matched Layers (PML) have been designed. PMLs have opened a large number of research directions and they are probably the most routinely used methods for modelling unbounded domains in geophysics. But in some cases, they turn out to be unstable. This is the case for some elastic media. We are thus considering the development of absorbing boundary conditions for elastodynamic media, and in particular for Tilted Transverse Isotropic media, which are of high interest for geophysical applications.

- **Asymptotic modeling**.

  During the last 30 years, mathematicians have developed and justified approximate models with multiscale asymptotic analysis to deal with problems involving singularly perturbed geometry or problems with coefficients of different magnitude.

  Numerically, all these approximate models are of interest since they allow to mesh the computational domain without taking into account the small characteristic lengths. this techniques lead to a reduction of the computation burden. Unfortunately, these methods do not have penetrated the numerical community since most of the results have been obtained for the two dimensional Laplacian.

  The research activity of Magique 3D aims in extending this theory to three-dimensional challenging problems involving wave propagation phenomena. We address time harmonic and time dependent problems for acoustic waves, electromagnetic waves and elastodynamic wave which is a very important topic for industry. Moreover, it remains numerous open questions in the underlying mathematical problems.

  Another important issue is the modeling of boundary layers which are not governed by the same model than the rest of the computational domain. It is rather challenging to derive and to justify some matching condition between the boundary layer and the rest of the physical domain for such multiphysical problems.

  More precisely, we have worked in 2014 on the following topics:

  – Eddy current modeling in the context of electrothermic applications for the design of electromagnetic devices, in collaboration with laboratories Ampère, Laplace, Inria Team MC2, IRMAR, and F.R.S.-FNRS;
  – Multiphysic asymptotic modeling of multi perforate plates in turbo reactors in collaboration with Onera.
  – Modeling of small heterogeneities for the three dimensional time domain wave equation. This reduced models is a generalization of the so called Lax-Foldy reduced model.
  – Modeling the propagation of ultrashort laser pulses in optical fibers.

3.3. High Performance methods for solving wave equations

Seismic Imaging of realistic 3D complex elastodynamic media does not only require advanced mathematical methods but also High Performing Computing (HPC) technologies, both from a software and hardware point of view. In the framework of our collaboration with Total, we are optimizing our algorithms, based on Discontinuous Galerkin methods, in the following directions.

- **Minimizing the communications between each processor.** One of the main advantages of Discontinuous Galerkin methods is that most of the calculi can be performed locally on each element of the mesh. The communications are carried out by the computations of fluxes on the faces of the

elements. Hence, there are only communications between elements sharing a common face. This represents a considerable gain compared with Continuous Finite Element methods where the communications have to be done between elements sharing a common degree of freedom. However, the communications can still be minimized by judiciously choosing the quantities to be passed from one element to another.

- **Hybrid MPI and OpenMP parallel programming.** Since the communications are one of the main bottlenecks for the implementation of the Discontinuous Galerkin in an HPC framework, it is necessary to avoid these communications between two processors sharing the same RAM. To achieve this aim, the partition of the mesh is not performed at the core level but at the chip level and the parallelization between two cores of the same chip is done using OpenMP while the parallelization between two cores of two different chips is done using MPI.
- **Porting the code on new architectures.** The goal is to test popular HPC architectures in the context of seismic imaging simulations. Current work concerns the new Intel Many Integrated Core Architecture (Intel MIC) of the Intel Xeon Phi co-processors and the upcoming stand-alone Intel processors.
- **Using Runtimes Systems.** One of the main issue of optimization of parallel code is the portability between different architectures. Indeed, many optimizations performed for a specific architecture are often useless for another architecture. In some cases, they may even reduce the performance of the code. One way to overcome this problem is to use task-based programming models through runtime libraries as StarPU (http://runtime.bordeaux.inria.fr/StarPU/ or PaRSEC (http://icl.cs.utk.edu/parsec/). However, until now, they have been mostly employed for solving linear algebra problems and our goal is to test their performances on realistic wave propagation simulations. This work is done in the framework of a collaboration with Inria Team Hiepacs and Georges Bosilca (University of Tennessee).

We are confident in the fact that the optimizations of the code will allow us to perform large-scale calculations and inversion of geophysical data for models and distributed data volumes with a resolution level impossible to reach in the past.

MAGNOME Project-Team

3. Research Program

3.1. Overview

Fundamental questions in the life sciences can now be addressed at an unprecedented scale through the combination of high-throughput experimental techniques and advanced computational methods from the computer sciences. The new field of *computational biology* or *bioinformatics* has grown around intense collaboration between biologists and computer scientists working towards understanding living organisms as *systems*. One of the key challenges in this study of systems biology is understanding how the static information recorded in the genome is interpreted to become dynamic systems of cooperating and competing biomolecules.

MAGNOME addresses this challenge through the development of informatic techniques for understanding the structure and history of eukaryote genomes: algorithms for genome analysis, data models for knowledge representation, stochastic hierarchical models for behavior of complex systems, and data mining and classification. Our work is in methods and algorithms for:

- **Genome annotation** for complete genomes, performing *syntactic* analyses to identify genes, and *semantic* analyses to map biological meaning to groups of genes [22], [5], [9], [10], [32], [33].
- **Integration of heterogenous data**, to build complete knowledge bases for storing and mining information from various sources, and for unambiguously exchanging this information between knowledge bases [1], [3], [25], [27], [21].
- **Ancestor reconstruction** using optimization techniques, to provide plausible scenarios of the history of genome evolution [10], [7], [28], [34].
- **Classification and logical inference**, to reliably identify similarities between groups of genetic elements, and infer rules through deduction and induction [8], [6], [9].
- **Hierarchical and comparative modeling**, to build mathematical models of the behavior of complex biological systems, in particular through combination, reutilization, and specialization of existing continuous and discrete models [24], [20][12].

The hundred- to thousand-fold decrease in sequencing costs seen in the past few years presents significant challenges for data management and large-scale data mining. MAGNOME's methods specifically address "scaling out," where resources are added by installing additional computation nodes, rather than by adding more resources to existing hardware. Scaling out adds capacity and redundancy to the resource, and thus fault tolerance, by enforcing data redundancy between nodes, and by reassigning computations to existing nodes as needed.

3.2. Comparative genomics

The central dogma of evolutionary biology postulates that contemporary genomes evolved from a common ancestral genome, but the large scale study of their evolutionary relationships is frustrated by the unavailability of these ancestral organisms that have long disappeared. However, this common inheritance allows us to discover these relationships through *comparison*, to identify those traits that are common and those that are novel inventions since the divergence of different lineages.

We develop efficient methodologies and software for associating biological information with complete genome sequences, in the particular case where several phylogenetically-related eukaryote genomes are studied simultaneously.

The methods designed by MAGNOME for comparative genome annotation, structured genome comparison, and construction of integrated models are applied on a large scale to: eukaryotes from the hemiascomycete class of yeasts [32], [33], [5], [9], [2], [10] and to lactic bacteria used in winemaking [31], [26]

3.3. Comparative modeling

A general goal of systems biology is to acquire a detailed quantitative understanding of the dynamics of living systems. Different formalisms and simulation techniques are currently used to construct numerical representations of biological systems, and a recurring challenge is that hand-tuned, accurate models tend to be so focused in scope that it is difficult to repurpose them. We claim that, instead of modeling individual processes *de novo*, a sustainable effort in building efficient behavioral models must proceed incrementally. *Hierarchical modeling* is one way of combining specific models into networks. Effective use of hierarchical models requires both formal definition of the semantics of such composition, and efficient simulation tools for exploring the large space of complex behaviors. We have combined uses theoretical results from formal methods and practical considerations from modeling applications to define a framework in which discrete and continuous models can communicate with a clear semantics. Hierarchical models can be assembled from existing models, and translated into their execution semantics and then simulated at multiple resolutions through multi-scale stochastic simulation. These models are compiled into a discrete event formalism capable of capturing discrete, continuous, stochastic, non deterministic and timed behaviors in an integrated and non-ambiguous way. Our long-term goal to develop a methodology in which we can **assemble a model** for a species of interest using a library of reusable models and a organism-level "schematic" determined by comparative genomics.

Comparative modeling is also a matter of reconciling experimental data with models [4] [20] and inferring new models through a combination of comparative genomics and successive refinement [29], [30].

MAMBA Team

3. Research Program

3.1. Introduction

At small spatial scales, or at spatial scales of individual matter components, where heterogeneities in the medium occur, agent-based models are developed ([0], [57], Dirk Drasdo's former associate team QUANTISS). Another approach, that is considered in the project-team MAMBA consists in considering gene expression at the individual level by stochastic processes [0] or by ordinary differential equations [0], or by a mixed representation of Markov processes and ordinary differential equations [0], the outputs of which quantify focused aspects of biological variability in a population of individuals (cells) under study.

Both these approaches complement the partial differential equation models considered on scales at which averages over the individual components behave sufficiently smoothly. Investigating the links between these models through scales is also part of our research [0]. Moreover, in order to quantitatively assess the adequacy between the biological phenomena we study and the mathematical models we use, we also develop inverse problem methods.

3.2. PDE analysis and simulation

PDEs arise at several levels of our models. Parabolic equations can be used for large cell populations and also for intracellular spatio-temporal dynamics of proteins and their messenger RNAs in gene regulatory networks, transport equations are used for protein aggregation / fragmentation models and for the cell division cycle in age-structured models of proliferating cell populations. Existence, uniqueness and asymptotic behaviour of solutions have been studied [0] [51] [50]. Other equations, of the integro-differential type, dedicated to describing the Darwinian evolution of a cell population according to a phenotypic trait, allowing exchanges with the environment, genetic mutations and reversible epigenetic modifications, are also used [59], [60], [58] [40]. Through multiscale analysis, they can be related to stochastic and free boundary models used in cancer modelling.

3.3. Inverse problems

When studying biological populations (usually cells or big molecules) using PDE models, identification of the functions and parameters that govern the dynamics of a model may be achieved to a certain extent by statistics performed on individuals to reconstruct the probability distribution of their relevant characteristics in the population they constitute, but quantitative observations at the individual level (e.g., fluorescence in single cells [4]) require sophisticated techniques and are most often difficult to obtain. Relying on the accuracy of a PDE model to describe the population dynamics, inverse problem methods offer a tractable alternative in model identification, and they are presently an active theme of research in MAMBA.

[0] Drasdo, Hoehme, Block, *J. Stat. Phys.*, 2007

[0] as in M. Sturrock et al., Spatial stochastic modelling of the Hes1 gene regulatory network: intrinsic noise can explain heterogeneity in embryonic stem cell differentiation, *Journal of The Royal Society Interface*, 2013

[0] as in A. Friedman et al, Asymptotic limit in a cell differentiation model with consideration of transcription, *J. Diff. Eq.*, 2012

[0] as in R. Yvinec et al., Adiabatic reduction of stochastic gene expression with jump Markov processes, *J. Math. Biol.*, 2013.

[0] H. Byrne and D. Drasdo, Individual-based and continuum models of growing cell populations: a comparison, *J. Math. Biol*, 2009

[0] B. Perthame, Transport equations in biology, Springer, 2007

3.4. Stochastic models

The link between stochastic processes and kinetic equations is a domain already present in our research [0] [53] and that we aim at developing further. They can be viewed either as complementary approaches, useful to take into account different scales (smaller scales for stochastic models, larger scales for mean-field limits), or even as two different viewpoints on the same problem [52], enriching each other. Neuroscience is a domain where this is particularly true because noise contributes significantly to the activity of neurons; this is particularly true for networks where mean field limits are derived from stochastic individual-based models and lead to fundamental questions on well-posedness and behaviours of the system [0]. One strength and originality of our project is our close connections and collaborations not only with probability theorists but also with statisticians.

3.5. Agent-based models

Agent-based systems consider each component individually. For example, in agent-based systems of multi-cellular systems the basic modeling unit is the cell, and each cell is considered [54]. This approach has advantages if the population of cells reveals inhomogeneities on small spatial scales as it occurs if organ architecture is represented [57], or if the number of cells in a particular state is small. Different approaches have been used to model cellular agents in multi-cellular systems in space, roughly distinguished by lattice models (e.g. [62]) and lattice-free (or off-lattice) models, in which the position of the cell can change gradually (e.g. [54], [56]). The dynamics of cells in lattice-based models is usually described by rules chosen to mimic the behavior of a cell including its physical behavior. The advantage of this approach is that it is simpler and simulation times for a given number of cells are shorter than in lattice-free models. In contrast, most lattice-free models attempt to parameterize cells by measurable values with a direct physical or biological meaning hence permitting identification of physiologically meaningful parameter ranges. This improves model simulation feasibility, since simulated parameter sensitivity analyses shows significant improvements when a high dimensional parameter space can be reduced. It also facilitates the development of systematic systems biology and systems medicine strategies to identify mechanisms underlying complex tissue organization processes ([7]). Moreover, it is straightforward to include relevant signal transduction and metabolic pathways in each cell within the framework of agent-based models, which is a key advantage in the present times where the interplay of components at many levels is more and more precisely studied [19].

[0] H. Byrne and D. Drasdo, Individual-based and continuum models of growing cell populations: a comparison, *J. Math. Biol*, 2009
[0] Cáceres, Carrillo, Perthame *J. Math. Neurosci.* 2011; Pakdaman, Perthame, Salort *Nonlinearity* 2010

MASAIE Project-Team

3. Research Program

3.1. Description

Our conceptual framework is that of Control Theory: the system is described by state variables with inputs (actions on the system) and outputs (the available measurements). Our system is either an epidemiological or immunological system or a harvested fish population. The control theory approach begins with the mathematical modeling of the system. When a “satisfying" model is obtained, this model is studied to understand the system. By “satisfying", an ambiguous word, we mean validation of the model. This depends on the objectives of the design of the model: explicative model, predictive model, comprehension model, checking hypotheses model. Moreover the process of modeling is not sequential. During elaboration of the model, a mathematical analysis is often done in parallel to describe the behavior of the proposed model. By behavior we intend not only asymptotic behavior but also such properties as observability, identifiability, robustness...

3.2. Structure and modeling

Problems in epidemiology, immunology and virology can be expressed as standard problems in control theory. But interesting new questions do arise. The control theory paradigm, input-output systems built out of simpler components that are interconnected, appears naturally in this context. Decomposing the system into several sub-systems, each of which endowed with certain qualitative properties, allows the behavior of the complete system to be deduced from the behavior of its parts. This paradigm, the toolbox of feedback interconnection of systems, has been used in the so-called theory of large-scale dynamic systems in control theory [23]. Reasons for decomposing are multiple. One reason is conceptual. For example connection of the immune system and the parasitic systems is a natural biological decomposition. Others reasons are for the sake of reducing algorithmic complexities or introducing intended behavior.... In this case subsystems may not have biological interpretation. For example a chain of compartments can be introduced to simulate a continuous delay [19], [20]. Analysis of the structure of epidemiological and immunological systems is vital because of the paucity of data and the dependence of behavior on biological hypotheses. The issue is to identify those parts of models that have most effects on dynamics. The concepts and techniques of interconnection of systems (large-scale systems) is useful in this regard.

In mathematical modeling in epidemiology and immunology, as in most other areas of mathematical modeling, there is always a trade-off between simple models, that omit details and are designed to highlight general qualitative behavior, and detailed models, usually designed for specific situations, including short-terms quantitative predictions. Detailed models are generally difficult to study analytically and hence their usefulness for theoretical purposes is limited, although their strategic value may be high. Simple models can be considered as building blocks of models that include detailed structure. The control theory tools of large-scale systems and interconnections of systems are a mean to conciliate the two approaches, simple models versus detailed systems.

3.3. Dynamic Problems

Many dynamical questions addressed by systems theory are precisely what biologist are asking. One fundamental problem is the problem of equilibria and their stability. To quote J.A. Jacquez

> A major project in deterministic modeling of heterogeneous populations is to find conditions for local and global stability and to work out the relations among these stability conditions, the threshold for epidemic take-off, and endemicity, and the basic reproduction number.

The basic reproduction number $\mathcal{R}_0$ is an important quantity in the study in epidemics. It is defined as the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible. The basic reproduction number $\mathcal{R}_0$ is often considered as the threshold quantity that determines when an infection can invade and persist in a new host population. To the problem of stability is related the problem of robustness, a concept from control theory. In other words how near is the system to an unstable one? Robustness is also in relation with uncertainty of the systems. This is a key point in epidemiological and immunological systems, since there are many sources of uncertainties in these models. The model is uncertain (parameters, functions, structure in some cases), the inputs also are uncertain and the outputs highly variable. That robustness is a fundamental issue and can be seen by means of an example: if policies in public health are to be taken from modeling, they must be based on robust reasons!

3.4. Observers

The concept of observer originates in control theory. This is particularly pertinent for epidemiological systems. To an input-output system, is associated the problem of reconstruction of the state. Indeed for a given system, not all the states are known or measured, this is particularly true for biological systems. This fact is due to a lot of reasons: this is not feasible without destroying the system, this is too expensive, there are no available sensors, measures are too noisy...The problem of knowledge of the state at present time is then posed. An observer is another system, whose inputs are the inputs and the outputs of the original system and whose output gives an estimation of the state of the original system at present time. Usually the estimation is required to be exponential. In other words an observer, using the signal information of the original system, reconstructs dynamically the state. More precisely, consider an input-output nonlinear system described by

$$\begin{cases} \dot{x} = f(x,u) \\ y = h(x), \end{cases} \tag{79}$$

where $x(t) \in \mathbb{R}^n$ is the state of the system at time t, $u(t) \in U \subset \mathbb{R}^m$ is the input and $y(t) \in \mathbb{R}^q$ is the measurable output of the system.
An observer for the system (1) is a dynamical system

$$\dot{\widehat{x}}(t) = g(\widehat{x}(t), y(t), u(t)), \tag{80}$$

where the map g has to be constructed such that: the solutions $x(t)$ and $\widehat{x}(t)$ of (1) and (2) satisfy for any initial conditions $x(0)$ and $\widehat{x}(0)$

$$\|x(t) - \widehat{x}(t)\| \leq c \, \|x(0) - \widehat{x}(0)\| \, e^{-at} \ , \ \forall t > 0$$

or at least $\|x(t) - \widehat{x}(t)\|$ converges to zero as time goes to infinity.
The problem of observers is completely solved for linear time-invariant systems (LTI). This is a difficult problem for nonlinear systems and is currently an active subject of research. The problem of observation and observers (software sensors) is central in nonlinear control theory. Considerable progress has been made in the last decade, especially by the “French school", which has given important contributions (J.P. Gauthier, H. Hammouri, E. Busvelle, M. Fliess, L. Praly, J.L. Gouzé, O. Bernard, G. Sallet) and is still very active in this area. Now the problem is to identify relevant classes of systems for which reasonable and computable observers can be designed. The concept of observer has been ignored by the modeler community in epidemiology, immunology and virology. To our knowledge one of the first use of an observer in virology was the work of Velasco-Hernandez J., Garcia J. and Kirschner D. [24] in modeling the chemotherapy of HIV, but this observer, based on classical linearization method, is a local observer and does not allow to deal with the nonlinearities.

3.5. Delays

Another crucial issue for biological systems is the question of delays. Delays, in control theory, are traditionally discrete (more exactly, the delays are lags) whereas in biology they usually are continuous and distributed. For example, the entry of a parasite into a cell initiates a cascade of events that ultimately leads to the production of new parasites. Even in a homogeneous population of cells, it is unreasonable to expect that the time to complete all these processes is the same for every cell. If we furthermore consider differences in cell activation state, metabolism, position in the cell cycle, pre-existing stores of nucleotides and other precursors needed for the reproduction of parasites, along with genetic variations in the parasite population, such variations in infection delay times become a near certainty. The rationale for studying continuous delays is supported by such considerations. In the literature on dynamical systems, we find a wealth of theorems dealing with delay differential equations. However they are difficult to apply. Control theory approache (interconnections of systems) is a mean to study the influence of continuous delays on the stability of such systems. We have obtained some results in this direction [6].

MNEMOSYNE Project-Team

3. Research Program

3.1. Integrative and Cognitive Neuroscience

The human brain is often considered as the most complex system dedicated to information processing. This multi-scale complexity, described from the metabolic to the network level, is particularly studied in integrative neuroscience, the goal of which is to explain how cognitive functions (ranging from sensorimotor coordination to executive functions) emerge from (are the result of the interaction of) distributed and adaptive computations of processing units, displayed along neural structures and information flows. Indeed, beyond the astounding complexity reported in physiological studies, integrative neuroscience aims at extracting, in simplifying models, regularities in space and functional mechanisms in time. From a spatial point of view, most neuronal structures (and particularly some of primary importance like the cortex, cerebellum, striatum, hippocampus) can be described through a regular organization of information flows and homogenous learning rules, whatever the nature of the processed information. From a temporal point of view, the arrangement in space of neuronal structures within the cerebral architecture also obeys a functional logic, the sketch of which is captured in models describing the main information flows in the brain, the corresponding loops built in interaction with the external and internal (bodily and hormonal) world and the developmental steps leading to the acquisition of elementary sensorimotor skills up to the most complex executive functions.

Three important characteristics are worth mentioning concerning these loops. Firstly, each of them sets a closed relation between the central nervous system and the rest of the world. This includes the external world (possibly including other intelligent agents), but also the internal world, with hormonal, physiological and bodily dimensions. Secondly, each of these loops can be described as a loop relating sensations to actions, in the wide sense of these terms: effectively, action can refer to acting in the real world, but also to modifying physiological parameters or controling neuronal activation. These loops have different constants of time, from immediate reflexes and sensorimotor adjustments to long term selection of motivation for action, the latter depending on hormonal and social parameters. Thirdly, each of the loops performs a learning reinforced by a primary (physiologically significant) or pseudo reward (sub-goal to be learned). As an illustration, we can mention respondent conditioning detecting stimuli anticipatory of primary rewards, episodic learning detecting multimodal events, and also more local phenomena like self-organization of topological structures. The gradual establishment of these loops and their mutual interactions give an interpretation of the resulting cognitive architecture as a synergetic system of memories.

In summary, integrative neuroscience builds, on an overwhelming quantity of data, a simplifying and interpretative grid suggesting homogenous local computations and a structured and logical plan for the development of cognitive functions. They arise from interactions and information exchange between neuronal structures and the external and internal world and also within the network of structures.

This domain is today very active and stimulating because it proposes, of course at the price of simplifications, global views of cerebral functioning and more local hypotheses on the role of subsets of neuronal structures in cognition. In the global approaches, the integration of data from experimental psychology and clinical studies leads to an overview of the brain as a set of interacting memories, each devoted to a specific kind of information processing [42]. It results also in longstanding and very ambitious studies for the design of cognitive architectures aiming at embracing the whole cognition. With the notable exception of works initiated by [38], most of these frameworks (e.g. Soar, ACT-R), though sometimes justified on biological grounds, do not go up to a *connectionist* neuronal implementation. Furthermore, because of the complexity of the resulting frameworks, they are restricted to simple symbolic interfaces with the internal and external world and to (relatively) small-sized internal structures. Our main research objective is undoubtly to build such a general purpose cognitive architecture (to model the brain *as a whole* in a systemic way), using a connectionist implementation and able to cope with a realistic environment.

3.2. Computational Neuroscience

From a general point of view, computational neuroscience can be defined as the development of methods from computer science and applied mathematics, to explore more technically and theoretically the relations between structures and functions in the brain [44], [32]. During the recent years this domain has gained an increasing interest in neuroscience and has become an essential tool for scientific developments in most fields in neuroscience, from the molecule to the system. In this view, all the objectives of our team can be described as possible progresses in computational neuroscience. Accordingly, it can be underlined that the systemic view that we promote can offer original contributions in the sense that, whereas most classical models in computational neuroscience focus on the better understanding of the structure/function relationship for isolated specific structures, we aim at exploring synergies between structures. Consequently, we target interfaces and interplay between heterogenous modes of computing, which is rarely addressed in classical computational neuroscience.

We also insist on another aspect of computational neuroscience which is, in our opinion, at the core of the involvement of computer scientists and mathematicians in the domain and on which we think we could particularly contribute. Indeed, we think that our primary abilities in numerical sciences imply that our developments are characterized above all by the effectiveness of the corresponding computations: We provide biologically inspired architectures with effective computational properties, such as robustness to noise, self-organization, on-line learning. We more generally underline the requirement that our models must also mimick biology through its most general law of homeostasis and self-adaptability in an unknown and changing environment. This means that we propose to numerically experiment such models and thus provide effective methods to falsify them.

Here, computational neuroscience means mimicking original computations made by the neuronal substratum and mastering their corresponding properties: computations are distributed and adaptive; they are performed without an homonculus or any central clock. Numerical schemes developed for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [29], [37] and were the basis for several contributions in our group [43], [40], [45]. Ensuring such a rigor in the computations associated to our systemic and large scale approach is of central importance.

Equally important is the choice for the formalism of computation, extensively discussed in the connectionist domain. Spiking neurons are today widely recognized of central interest to study synchronization mechanisms and neuronal coupling at the microscopic level [30]; the associated formalism [35] can be possibly considered for local studies or for relating our results with this important domain in connectionism. Nevertheless, we remain mainly at the mesoscopic level of modeling, the level of the neuronal population, and consequently interested in the formalism developed for dynamic neural fields [27], that demonstrated a richness of behavior [31] adapted to the kind of phenomena we wish to manipulate at this level of description. Our group has a long experience in the study and adaptation of the properties of neural fields [40], [41] and their use for observing the emergence of typical cortical properties [34]. In the envisioned development of more complex architectures and interplay between structures, the exploration of mathematical properties such as stability and boundedness and the observation of emerging phenomena is one important objective. This objective is also associated with that of capitalizing our experience and promoting good practices in our software production (*cf.* § 5.1). In summary, we think that this systemic approach also brings to computational neuroscience new case studies where heterogenous and adaptive models with various time scales and parameters have to be considered jointly to obtain a mastered substratum of computation. This is particularly critical for large scale deployments, as we will discuss in § 5.1).

3.3. Machine Learning

The adaptive properties of the nervous system are certainly among its most fascinating characteristics, with a high impact on our cognitive functions. Accordingly, machine learning is a domain [36] that aims at giving such characteristics to artificial systems, using a mathematical framework (probabilities, statistics, data analysis, etc.). Some of its most famous algorithms are directly inspired from neuroscience, at different levels.

Connectionist learning algorithms implement, in various neuronal architectures, weight update rules, generally derived from the hebbian rule, performing non supervised (e.g. Kohonen self-organizing maps), supervised (e.g. layered perceptrons) or associative (e.g. Hopfield recurrent network) learning. Other algorithms, not necessarily connectionist, perform other kinds of learning, like reinforcement learning. Machine learning is a very mature domain today and all these algorithms have been extensively studied, at both the theoretical and practical levels, with much success. They have also been related to many functions (in the living and artificial domains) like discrimination, categorisation, sensorimotor coordination, planning, etc. and several neuronal structures have been proposed as the substratum for these kinds of learning [33], [26]. Nevertheless, we believe that, as for previous models, machine learning algorithms remain isolated tools, whereas our systemic approach can bring original views on these problems.

At the cognitive level, most of the problems we face do not rely on only one kind of learning and require instead skills that have to be learned in preliminary steps. That is the reason why cognitive architectures are often referred to as systems of memory, communicating and sharing information for problem solving. Instead of the classical view in machine learning of a flat architecture, a more complex network of modules must be considered here, as it is the case in the domain of deep learning. In addition, our systemic approach brings the question of incrementally building such a system, with a clear inspiration from developmental sciences. In this perspective, modules can generate internal signals corresponding to internal goals, predictions, error signals, able to supervise the learning of other modules (possibly endowed with a different learning rule), supposed to become autonomous after an instructing period. A typical example is that of episodic learning (in the hippocampus), storing declarative memory about a collection of past episods and supervising the training of a procedural memory in the cortex.

At the behavioral level, as mentionned above, our systemic approach underlines the fundamental links between the adaptive system and the internal and external world. The internal world includes proprioception and interoception, giving information about the body and its needs for integrity and other fundamental programs. The external world includes physical laws that have to be learned and possibly intelligent agents for more complex interactions. Both involve sensors and actuators that are the interfaces with these worlds and close the loops. Within this rich picture, machine learning generally selects one situation that defines useful sensors and actuators and a corpus with properly segmented data and time, and builds a specific architecture and its corresponding criteria to be satisfied. In our approach however, the first question to be raised is to discover what is the goal, where attention must be focused on and which previous skills must be exploited, with the help of a dynamic architecture and possibly other partners. In this domain, the behavioral and the developmental sciences, observing how and along which stages an agent learns, are of great help to bring some structure to this high dimensional problem.

At the implementation level, this analysis opens many fundamental challenges, hardly considered in machine learning : stability must be preserved despite on-line continuous learning; criteria to be satisfied often refer to behavioral and global measurements but they must be translated to control the local circuit level; in an incremental or developmental approach, how will the development of new functions preserve the integrity and stability of others? In addition, this continous re-arrangement is supposed to involve several kinds of learning, at different time scales (from msec to years in humans) and to interfer with other phenomena like variability and meta-plasticity.

In summary, our main objective in machine learning is to propose on-line learning systems, where several modes of learning have to collaborate and where the protocoles of training are realistic. We promote here a *really autonomous* learning, where the agent must select by itself internal resources (and build them if not available) to evolve at the best in an unknown world, without the help of any *deus-ex-machina* to define parameters, build corpus and define training sessions, as it is generally the case in machine learning. To that end, autonomous robotics (*cf.* § 3.4) is a perfect testbed.

3.4. Autonomous Robotics

Autonomous robots are not only convenient platforms to implement our algorithms; the choice of such platforms is also motivated by theories in cognitive science and neuroscience indicating that cognition emerges

from interactions of the body in direct loops with the world and develops interesting specificities accordingly. For example, internal representations can be minimized (opposite to building complex and hierarchical representations) and compensated by more simple strategies [28], more directly coupling perception and action and more efficient to react quickly in the changing environment (for example, instead of memorizing details of an object, just memorizing the eye movement to foveate it: the world itself is considered as an external memory). In this view for the *embodiment of cognition*, learning is intrinsically linked to sensorimotor loops and to a real body interacting with a real environment.

A real autonomy can be obtained only if the robot is able to define its goal by itself, without the specification of any high level and abstract cost function or rewarding state. To ensure such a capability, we propose to endow the robot with an artificial physiology, corresponding to perceive some kind of pain and pleasure. It may consequently discriminate internal and external goals (or situations to be avoided). This will mimick circuits related to fundamental needs (e.g. hunger and thirst) and to the preservation of bodily integrity. An important objective is to show that more abstract planning capabilities can arise from these basic goals.

A real autonomy with an on-line continuous learning as described in § 3.3 will be made possible by the elaboration of protocols of learning, as it is the case, in animal conditioning, for experimental studies where performance on a task can be obtained only after a shaping in increasingly complex tasks. Similarly, developmental sciences can teach us about the ordered elaboration of skills and their association in more complex schemes. An important challenge here is to translate these hints at the level of the cerebral architecture.

As a whole, autonomous robotics permits to assess the consistency of our models in realistic condition of use and offers to our colleagues in behavioral sciences an object of study and comparison, regarding behavioral dynamics emerging from interactions with the environment, also observable at the neuronal level.

In summary, our main contribution in autonomous robotics is to make autonomy possible, by various means corresponding to endow robots with an artificial physiology, to give instructions in a natural and incremental way and to prioritize the synergy between reactive and robust schemes over complex planning structures.

MODEMIC Project-Team

3. Research Program

3.1. Modeling and simulating microbial ecosystems

The chemostat model is quite popular in microbiology and bioprocess engineering [60], [62]. Although the wording "chemostat" refers to the experimental apparatus dedicated to continuous culture, invented in the fifties by Monod and Novick & Szilard, the chemostat model often serves as a mathematical representation of biotic/abiotic interactions in more general (industrial or natural) frameworks of microbial ecology. The team carries a significant activity about generalizations and extensions of the classical model (see Equation (1) and Section 3.1.1) which assumes that the sizes of the populations are large and that the biomass can be faithfully represented as a set of deterministic continuous variables.

However recent observations tools based notably on molecular biology (e.g. molecular fingerprints) allow to distinguish much more precisely than in the past the internal composition of biomass. In particular, it has been reported by biologists that minority species could play an important role during transients (in the initialization phase of bio-processes or when the ecosystem is recovering from disturbances), that cannot be satisfactorily explained by the above deterministic models because the size of those populations could be too small for these models to be valid.

Therefore, we are studying extension of the classical model that could integrate stochastic/continuous macroscopic aspects, or microscopic/discrete aspects (in terms of population size or even with explicit individually based representation of the bacteria), as well as hybrid representations. One important question is the links between these chemostat models (see Section 3.1.2).

3.1.1. About the chemostat model

The classical mathematical chemostat model:

$$\begin{aligned} \dot{s} &= -\sum_{j=1}^{n} \frac{1}{y_j} \mu_j(s)\, x_j + D\,(s_{in} - s) \\ \dot{x}_i &= \mu_i(s)\, x_i - D\, x_i \qquad (i = 1 \cdots n) \end{aligned} \tag{81}$$

for n species in concentrations x_i competing for a substrat in concentration s, leads to the so-called "Competitive Exclusion Principle", that states that generically no more species than limiting resources can survive on a long term [61]. Apart some very precise laboratory experiments that have validated this principle, such an exclusion is rarely observed in practice.

Several possible improvements of the model (1) need to be investigated, related to biologists' knowledge and observations, in order to provide better interpretations and predictive tools. Various extensions have already been studied in the literature (e.g. crowding effect, inter-specific interactions, predating, spatialization, time-varying inputs...) to which the team has also contributed. This is always an active research topic in bio-mathematics and theoretical ecology, and several questions remains open or unclear, although numerical simulations guide the results to be proven.

Thanks to the proximity with biologists, the team is in position to propose new extensions relevant for experiments or processes conducted among the application partners. Among them, we can mention: intra and inter-specific interactions terms between microbial species; distinction between planktonic and attached biomass; effects of interconnected vessels; consideration of maintenance or variable yield in the growth reactions; coupling with membrane fouling mechanisms.

Our philosophy is to study how complex or not very well known mechanisms could be represented satisfactorily by simple models. It often happens that these mechanisms have different time scales (for instance the flocculation of bacteria is expected to be much faster than the biomass growth), and we typically use singular perturbations techniques to produce reduced models.

3.1.2. Stochastic and multi-scale models

Comparatively to deterministic differential equations models, quite few stochastic models of microbial growth have been worked out in the literature. Nonetheless, numerous problems could benefit from such an approach (dynamics with small population sizes, persistence and extinction, random environments...).

For example, the need to clarify the role of minority species conducts to revisit thoroughly the chemostat model at a microscopic level, with birth and death or pure jump processes, and to investigate which kind of continuous models it raises at a macroscopic scale. For this purpose, we consider the general framework of Markov processes [59].

It also happens that minority species cohabit with other populations of much larger size, or fluctuate with time between small and large sizes. There is consequently a need to build new "hybrid" models, that have individual-based and deterministic continuous parts at the same time. The persistence (temporarily or not) of minority species on the long term is quite a new questioning spread in several applications domains at the Inra Institute.

Continuous cultures of micro-organisms often face random abiotic environments, that could be considered as random switching between favorable or unfavorable environments. This feature could lead to non-intuitive behaviors in long run, concerning persistence or extinction of populations. We consider here the framework of piecewise deterministic Markov processes [58].

3.1.3. Computer simulation

The simulation of dynamical models of microbial ecosystems with the features described in Section 3.1.2 raises specific and original algorithmic problems:

- simultaneous presence in the same algorithms of both continuous variables (concentration of chemicals or very large populations) and discrete (when the population has a very small number of individuals),
- simultaneous presence in the same algorithms of stochastic aspects (for demographic and environmental noises) and deterministic ones (when the previous noises are negligible at macroscopic scales)
- use of individual-based models (IBM) (usually for small population sizes).

We believe that these questions must be addressed in a rigorous mathematical framework and that their solutions as efficient algorithms are a formidable scientific challenge.

3.2. Identification and control

3.2.1. Models identification and state estimation

Growth kinetics is usually one of the crucial ingredients in the modeling of microbial growth. Although the specific growth rate functions and their parameters can be identified in pure cultures (and can be estimated with accuracy in laboratory experiments), it is often an issue to extrapolate this knowledge in industrial setup or in mixed cultures. The parameters of these functions could change with their chemical and physical environment, and species interactions could inhibit or promote a strain that is expected to dominate or to be dominated in an multi-species ecosystem. Moreover, we need to estimate the state variables of the models.

We aim at developing effective tools for the on-line reconstruction of growth curves (and of their parameters) and/or state variables, along with the characteristics of microbial ecosystems:

- It is not always possible to drive a biological system for exploring a large subset of the state space, and open-loop dynamics could be unstable when far from locally stable equilibria (for instance under inhibition growth).
- The number of functional groups of species and the nature of their interactions (competition, mutualism, neutral) are not always known a priori and need to be estimated.

We look for observers or filters based methods (or alternatives), as well as estimation procedures, with the typical difficulty that for biological systems and their outputs it is rarely straightforward to write the models into a canonical observation form. However, our objective is to obtain an adjustable or guaranteed speed of convergence of the estimators.

3.2.2. Optimal design and control

For practitioners, an expected outcome of the models is to bring improvements in the design and real-time operation of the processes. This naturally leads to mathematical formulations of optimization, stabilizing control or optimal control problems. We distinguish two families of problems:

- *Process design and control within an industrial setup.* Typically one aims at obtaining small residence times for given input-output performances and (globally) stable processes. The design questions consist in studying on the models if particular interconnections and fill strategies allow to obtain significant gains. The specificity of the models and the inputs constraints can lead to systems that are not locally controllable, and thus the classical linearizing techniques do not work. This leaves open some problems for the determination of globally stabilizing feedback or optimal syntheses.
- *Design and control for resource preservation in natural environments (such as lakes, soil bio-remediation...).* Here, the spatial heterogeneity of the resource might be complex and/or not well known. We look for sparse spatial representations in order to apply finite dimensional tools of state-space systems.

In both cases, one faces model uncertainty and partial measurements that often require to couple the techniques developed in Section 3.2.1 .

MOISE Project-Team

3. Research Program

3.1. Introduction

Geophysical flows generally have a number of particularities that make it difficult to model them and that justify the development of specifically adapted mathematical and numerical methods:

- Geophysical flows are non-linear. There is often a strong interaction between the different scales of the flows, and small-scale effects (smaller than mesh size) have to be modelled in the equations.
- Every geophysical episode is unique: a field experiment cannot be reproduced. Therefore the validation of a model has to be carried out in several different situations, and the role of the data in this process is crucial.
- Geophysical fluids are non closed systems, i.e. there are always interactions between the different components of the environment (atmosphere, ocean, continental water, etc.). Boundary terms are thus of prime importance.
- Geophysical flows are often modeled with the goal of providing forecasts. This has several consequences, like the usefulness of providing corresponding error bars or the importance of designing efficient numerical algorithms to perform computations in a limited time.

Given these particularities, the overall objectives of the MOISE project-team described earlier will be addressed mainly by using the mathematical tools presented in the following.

3.2. Numerical Modelling

Models allow a global view of the dynamics, consistent in time and space on a wide spectrum of scales. They are based on fluid mechanics equations and are complex since they deal with the irregular shape of domains, and include a number of specific parameterizations (for example, to account for small-scale turbulence, boundary layers, or rheological effects). Another fundamental aspect of geophysical flows is the importance of non-linearities, i.e. the strong interactions between spatial and temporal scales, and the associated cascade of energy, which of course makes their modelling more complicated.

Since the behavior of a geophysical fluid generally depends on its interactions with others (e.g. interactions between ocean, continental water, atmosphere and ice for climate modelling), building a forecasting system often requires **coupling different models**. Several kinds of problems can be encountered, since the models to be coupled may differ in numerous respects: time and space resolution, physics, dimensions. Depending on the problem, different types of methods can be used, which are mainly based on open and absorbing boundary conditions, multi-grid theory, domain decomposition methods, and optimal control methods.

3.3. Data Assimilation and Inverse Methods

Despite their permanent improvement, models are always characterized by an imperfect physics and some poorly known parameters (e.g. initial and boundary conditions). This is why it is important to also have **observations** of natural systems. However, observations provide only a partial (and sometimes very indirect) view of reality, localized in time and space.

Since models and observations taken separately do not allow for a deterministic reconstruction of real geophysical flows, it is necessary to use these heterogeneous but complementary sources of information simultaneously, by using **data assimilation methods**. These tools for **inverse modelling** are based on the mathematical theories of optimal control and stochastic filtering. Their aim is to identify system parameters which are poorly known in order to correct, in an optimal manner, the model trajectory, bringing it closer to the available observations.

Variational methods are based on the minimization of a function measuring the discrepancy between a model solution and observations, using optimal control techniques for this purpose. The model inputs are then used as control variables. The Euler Lagrange condition for optimality is satisfied by the solution of the “Optimality System" (OS) that contains the adjoint model obtained by derivation and transposition of the direct model. It is important to point out that this OS contains all the available information: model, data and statistics. The OS can therefore be considered as a generalized model. The adjoint model is a very powerful tool which can also be used for other applications, such as sensitivity studies.

Stochastic filtering is the basic tool in the sequential approach to the problem of data assimilation into numerical models, especially in meteorology and oceanography. The (unknown) initial state of the system can be conveniently modeled by a random vector, and the error of the dynamical model can be taken into account by introducing a random noise term. The goal of filtering is to obtain a good approximation of the conditional expectation of the system state (and of its error covariance matrix) given the observed data. These data appear as the realizations of a random process related to the system state and contaminated by an observation noise.

The development of data assimilation methods in the context of geophysical fluids, however, is difficult for several reasons:

- the models are often strongly non-linear, whereas the theories result in optimal solutions only in the context of linear systems;
- the model error statistics are generally poorly known;
- the size of the model state variable is often quite large, which requires dealing with huge covariance matrices and working with very large control spaces;
- data assimilation methods generally increase the computational costs of the models by one or two orders of magnitude.

Such methods are now used operationally (after 15 years of research) in the main meteorological and oceanographic centers, but tremendous development is still needed to improve the quality of the identification, to reduce their cost, and to make them available for other types of applications.

A challenge of particular interest consists in developing methods for assimilating image data. Indeed, images and sequences of images represent a large amount of data which are currently underused in numerical forecast systems. However, despite their huge informative potential, images are only used in a qualitative way by forecasters, mainly because of the lack of an appropriate methodological framework.

3.4. Sensitivity Analysis - Quantification of Uncertainties

Due to the strong non-linearity of geophysical systems and to their chaotic behavior, the dependence of their solutions on external parameters is very complex. Understanding the relationship between model parameters and model solutions is a prerequisite to design better models as well as better parameter identification. Moreover, given the present strong development of forecast systems in geophysics, the ability to provide an estimate of the uncertainty of the forecast is of course a major issue. However, the systems under consideration are very complex, and providing such an estimation is very challenging. Several mathematical approaches are possible to address these issues, using either variational or stochastic tools.

Variational approach. In the variational framework, the sensitivity is the gradient of a response function with respect to the parameters or the inputs of the model. The adjoint techniques can therefore be used for such a purpose. If sensitivity is sought in the context of a forecasting system assimilating observations, the optimality system must be derived. This leads to the study of second-order properties: spectrum and eigenvectors of the Hessian are important information on system behavior.

Global stochastic approach. Using the variational approach to sensitivity leads to efficient computations of complex code derivatives. However, this approach to sensitivity remains local because derivatives are generally computed at specific points. The stochastic approach of uncertainty analysis aims at studying global criteria describing the global variabilities of the phenomena. For example, the Sobol sensitivity index is given by the ratio between the output variance conditionally to one input and the total output variance. The computation of such quantities leads to statistical problems. For example, the sensitivity indices have to be efficiently estimated from a few runs, using semi or non-parametric estimation techniques. The stochastic modeling of the input/output relationship is another solution.

MORPHEME Project-Team

3. Research Program

3.1. Research Program

The recent advent of an increasing number of new microscopy techniques giving access to high throughput screenings and micro or nano-metric resolutions provides a means for quantitative imaging of biological structures and phenomena. To conduct quantitative biological studies based on these new data, it is necessary to develop non-standard specific tools. This requires using a multi-disciplinary approach. We need biologists to define experiment protocols and interpret the results, but also physicists to model the sensors, computer scientists to develop algorithms and mathematicians to model the resulting information. These different expertises are combined within the Morpheme team. This generates a fecund frame for exchanging expertise, knowledge, leading to an optimal framework for the different tasks (imaging, image analysis, classification, modeling). We thus aim at providing adapted and robust tools required to describe, explain and model fundamental phenomena underlying the morphogenesis of cellular and supra-cellular biological structures. Combining experimental manipulations, in vivo imaging, image processing and computational modeling, we plan to provide methods for the quantitative analysis of the morphological changes that occur during development. This is of key importance as the morphology and topology of mesoscopic structures govern organ and cell function. Alterations in the genetic programs underlying cellular morphogenesis have been linked to a range of pathologies.
Biological questions we will focus on include:

1. what are the parameters and the factors controlling the establishment of ramified structures? (Are they really organize to ensure maximal coverage? How are genetical and physical constraints limiting their morphology?),
2. how are newly generated cells incorporated into reorganizing tissues during development? (is the relative position of cells governed by the lineage they belong to?)

Our goal is to characterize different populations or development conditions based on the shape of cellular and supra-cellular structures, e.g. micro-vascular networks, dendrite/axon networks, tissues from 2D, 2D+t, 3D or 3D+t images (obtained with confocal microscopy, video-microscopy, photon-microscopy or micro-tomography). We plan to extract shapes or quantitative parameters to characterize the morphometric properties of different samples. On the one hand, we will propose numerical and biological models explaining the temporal evolution of the sample, and on the other hand, we will statistically analyze shapes and complex structures to identify relevant markers for classification purposes. This should contribute to a better understanding of the development of normal tissues but also to a characterization at the supra-cellular scale of different pathologies such as Alzheimer, cancer, diabetes, or the Fragile X Syndrome. In this multidisciplinary context, several challenges have to be faced. The expertise of biologists concerning sample generation, as well as optimization of experimental protocols and imaging conditions, is of course crucial. However, the imaging protocols optimized for a qualitative analysis may be sub-optimal for quantitative biology. Second, sample imaging is only a first step, as we need to extract quantitative information. Achieving quantitative imaging remains an open issue in biology, and requires close interactions between biologists, computer scientists and applied mathematicians. On the one hand, experimental and imaging protocols should integrate constraints from the downstream computer-assisted analysis, yielding to a trade-off between qualitative optimized and quantitative optimized protocols. On the other hand, computer analysis should integrate constraints specific to the biological problem, from acquisition to quantitative information extraction. There is therefore a need of specificity for embedding precise biological information for a given task. Besides, a level of generality is also desirable for addressing data from different teams acquired with different protocols and/or sensors. The mathematical modeling of the physics of the acquisition system will yield higher performance reconstruction/restoration algorithms in terms of accuracy. Therefore, physicists and computer scientists have to work together. Quantitative information extraction also has to deal with both the complexity of the structures of interest (e.g., very

dense network, small structure detection in a volume, multiscale behavior, ...) and the unavoidable defects of in vivo imaging (artifacts, missing data, ...). Incorporating biological expertise in model-based segmentation methods provides the required specificity while robustness gained from a methodological analysis increases the generality. Finally, beyond image processing, we aim at quantifying and then statistically analyzing shapes and complex structures (e.g., neuronal or vascular networks), static or in evolution, taking into account variability. In this context, learning methods will be developed for determining (dis)similarity measures between two samples or for determining directly a classification rule using discriminative models, generative models, or hybrid models. Besides, some metrics for comparing, classifying and characterizing objects under study are necessary. We will construct such metrics for biological structures such as neuronal or vascular networks. Attention will be paid to computational cost and scalability of the developed algorithms: biological experimentations generally yield huge data sets resulting from high throughput screenings. The research of Morpheme will be developed along the following axes:

- **Imaging:** this includes i) definition of the studied populations (experimental conditions) and preparation of samples, ii) definition of relevant quantitative characteristics and optimized acquisition protocol (staining, imaging, ...) for the specific biological question, and iii) reconstruction/restoration of native data to improve the image readability and interpretation.
- **Feature extraction:** this consists in detecting and delineating the biological structures of interest from images. Embedding biological properties in the algorithms and models is a key issue. Two main challenges are the variability, both in shape and scale, of biological structures and the huge size of data sets. Following features along time will allow to address morphogenesis and structure development.
- **Classification/Interpretation:** considering a database of images containing different populations, we can infer the parameters associated with a given model on each dataset from which the biological structure under study has been extracted. We plan to define classification schemes for characterizing the different populations based either on the model parameters, or on some specific metric between the extracted structures.
- **Modeling:** two aspects will be considered. This first one consists in modeling biological phenomena such as axon growing or network topology in different contexts. One main advantage of our team is the possibility to use the image information for calibrating and/or validating the biological models. Calibration induces parameter inference as a main challenge. The second aspect consists in using a prior based on biological properties for extracting relevant information from images. Here again, combining biology and computer science expertise is a key point.

MYCENAE Project-Team

3. Research Program

3.1. Project team positioning

The main goal of MYCENAE is to address crucial questions arising from both Neuroendocrinology and Neuroscience from a mathematical perspective. The choice and subsequent study of appropriate mathematical formalisms to investigate these dynamics is at the core of MYCENAE's scientific foundations: slow-fast dynamical systems with multiple time scales, mean-field approaches subject to limit-size and stochastic effects, transport-like partial differential equations (PDE) and stochastic individual based models (SIBM).

The scientific positioning of MYCENAE is on the way between Mathematical Biology and Mathematics: we are involved both in the modeling of physiological processes and in the deep mathematical analysis of models, whether they be (i) models developed (or under development) within the team (ii) models developed by collaborating teams or (iii) benchmark models from the literature.

Our research program is grounded on previous results obtained in the framework of the REGATE (REgulation of the GonAdoTropE axis) Large Scale Initiative Action and the SISYPHE project team on the one hand, and the Mathematical Neuroscience Team in the Center for Interdisciplinary Research in Biology (Collège de France), on the other hand. Several of our research topics are related to the study and generalization of 2 master models: a 4D, multiscale in time, nonlinear model based on coupled FitzHugh-Nagumo dynamics that has proved to be a fruitful basis for the study of the complex oscillations in hypothalamic GnRH dynamics [38], [37], and a nD, multiscale in space, system of weakly-coupled non conservative transport equations that underlies our approach of gonadal cell dynamics [39],[6]. Most our topics in mathematical neuroscience deal with the study of complex oscillatory behaviors exhibited either by single neurons or as emergent macroscopic properties of neural networks, from both a deterministic and stochastic viewpoint.

3.2. Numerical and theoretical studies of slow-fast systems with complex oscillations

In dynamical systems with at least three state variables, the presence of different time scales favors the appearance of complex oscillatory solutions. In this context, with (at least) two slow variables MixedMode Oscillations (MMO) dynamics can arise. MMOs are small and large amplitude oscillations combined in a single time series. The last decade has witnessed a significant amount of research on this topic, including studies of folded singularities, construction of MMOs using folded singularities in combination with global dynamics, effects of additional time scales, onset of MMOs via singular Hopf bifurcations, as well as generalization to higher dimensions. In the same period, many applications to neuroscience emerged [7]. On the other hand, bursting oscillations, another prototype of complex oscillations can occur in systems with (at least) two fast variables. Bursting has been observed in many biological contexts, in particular in the dynamics of pancreatic cells, neurons, and other excitable cells. In neuronal dynamics a burst corresponds to a series of spikes, interspersed with periods of quiescent behavior, called inter-burst intervals. We are interested in systems combining bursting, MMOs and canards. One of the interesting directions is torus canards, which are canard-like structures occurring in systems combining canard explosion with fast rotation [3]. Torus canards help understand transitions from spiking or MMO dynamics to bursting. Another study on the boundary of bursting and MMOs is the work of [43] on the so-called plateau bursting. A major challenge in this direction is to gain a complete understanding of the transition from "3 time scales" to "2 fast/ 1 slow" (bursting) and then to "1 fast/ 2 slow (MMOs)". Also, a key challenge that we intend to tackle in the next few years is that of large dynamical systems with many fast and many slow variables, which additionally are changing in time and/or in phase space. We aim to pursue this research direction both at theoretical and computational level, using numerical continuation approaches based on the location of unstable trajectories by using fixed point methods, rather than simulation, to locate trajectories.

3.3. Non conservative transport equations for cell population dynamics

Models for physiologically-structured populations can be considered to derive from the so-called McKendrick-Von Foerster equation or renewal equation that has been applied and generalized in different applications of population dynamics, including ecology, epidemiology and cell biology. Renewal equations are PDE transport equations that are written so as to combine conservation laws (e.g. on the total number of individuals) with additional terms related to death or maturation, that blur the underlying overall balance law.

The development of ovarian follicles is a tightly-controlled physiological and morphogenetic process, that can be investigated from a middle-out approach starting at the cell level. To describe the terminal stages of follicular development on a cell kinetics basis and account for the selection process operated amongst follicles, we have developed a multiscale model describing the cell density in each follicle, that can be roughly considered as a system of weakly-coupled, non conservative transport equations with controlled velocities and source term. Even if, in some sense, this model belongs to the class of renewal equations for structured populations, it owns a number of specificities that render its theoretical and numerical analysis particularly challenging: 2 structuring variables (per follicle, leading as a whole to $2n$D system), control terms operating on the velocities and source term, and formulated from moments of the unknowns, discontinuities both in the velocities and density on internal boundaries of the domain representing the passage from one cell phase to another.

On the theoretical ground, the well-posedness (existence and uniqueness of weak solutions with bounded initial data) has been established in [10], while associated control problems have been studied in the framework of hybrid optimal control [4]. On the numerical ground, the formalism dedicated to the simulation of these hyperbolic-like PDEs is that of finite volume method. Part of the numerical strategy consists in combining in the most efficient way low resolution numerical schemes (such as the first-order Godunov scheme), that tend to be diffusive, with high resolution schemes (such as the Lax Wendroff second-order scheme), that may engender oscillations in the vicinity of discontinuities [2], with a critical choice of the limiter functions. The 2D finite volume schemes are combined with adaptive mesh refinement through a multi-resolution method [16] and implemented in a problem-specific way on parallel architecture [1].

3.4. Macroscopic limits of stochastic neural networks and neural fields

The coordinated activity of the cortex is the result of the interactions between a very large number of cells. Each cell is well described by a dynamical system, that receives non constant input which is the superposition of an external stimulus, noise and interactions with other cells. Most models describing the emergent behavior arising from the interaction of neurons in large-scale networks have relied on continuum limits ever since the seminal work of Wilson and Cowan and Amari [44], [36]. Such models tend to represent the activity of the network through a macroscopic variable, the population-averaged firing rate.

In order to rationally describe neural fields and more generally large cortical assemblies, one should yet base their approach on what is known of the microscopic neuronal dynamics. At this scale, the equation of the activity is a set of stochastic differential equations in interaction. Obtaining the equations of evolution of the effective mean-field from microscopic dynamics is a very complex problem which belongs to statistical physics. As in the case of the kinetic theory of gases, macroscopic states are defined by the limit of certain quantities as the network size tends to infinity. When such a limit theorem is proved, one can be ensured that large networks are well approximated by the obtained macroscopic system. Qualitative distinctions between the macroscopic limit and finite-sized networks (finite-size effects), occurs in such systems. We have been interested in the relevant mathematical approaches dealing with macroscopic limits of stochastic neuronal networks, that are expressed in the form of a complex integro-differential stochastic implicit equations of McKean-Vlasov type including a new mathematical object, the spatially chaotic Brownian motion [24].

The major question consists in establishing the fundamental laws of the collective behaviors cortical assemblies in a number of contexts motivated by neuroscience, such as communication delays between cells [12], [11] or spatially extended areas, which is the main topic of our current research. In that case additional difficulties arise, since the connection between different neurons, as well as delays in communications, depend on

space in a correlated way, leading to the singular dependence of the solutions in space, which is not measurable.

NEUROMATHCOMP Project-Team

3. Research Program

3.1. Neural networks dynamics

The study of neural networks is certainly motivated by the long term goal to understand how brain is working. But, beyond the comprehension of brain or even of simpler neural systems in less evolved animals, there is also the desire to exhibit general mechanisms or principles at work in the nervous system. One possible strategy is to propose mathematical models of neural activity, at different space and time scales, depending on the type of phenomena under consideration. However, beyond the mere proposal of new models, which can rapidly result in a plethora, there is also a need to understand some fundamental keys ruling the behaviour of neural networks, and, from this, to extract new ideas that can be tested in real experiments. Therefore, there is a need to make a thorough analysis of these models. An efficient approach, developed in our team, consists of analysing neural networks as dynamical systems. This allows to address several issues. A first, natural issue is to ask about the (generic) dynamics exhibited by the system when control parameters vary. This naturally leads to analyse the bifurcations occurring in the network and which phenomenological parameters control these bifurcations. Another issue concerns the interplay between neuron dynamics and synaptic network structure.

In this spirit, our team has been able to characterize the generic dynamics exhibited by models such as Integrate and Fire models [10], conductance-based Integrate and Fire models [53], [57], [45], models of epilepsy [81], effects of synaptic plasticity [77], [78], homeostasis and intrinsic plasticity [8].
Selected publications on this topic.

3.2. Mean-field approaches

Modeling neural activity at scales integrating the effect of thousands of neurons is of central importance for several reasons. First, most imaging techniques are not able to measure individual neuron activity (“microscopic” scale), but are instead measuring mesoscopic effects resulting from the activity of several hundreds to several hundreds of thousands of neurons. Second, anatomical data recorded in the cortex reveal the existence of structures, such as the cortical columns, with a diameter of about 50μm to 1mm, containing of the order of one hundred to one hundred thousand neurons belonging to a few different species. The description of this collective dynamics requires models which are different from individual neurons models. In particular, when the number of neurons is large enough averaging effects appear, and the collective dynamics is well described by an effective mean-field, summarizing the effect of the interactions of a neuron with the other neurons, and depending on a few effective control parameters. This vision, inherited from statistical physics requires that the space scale be large enough to include a large number of microscopic components (here neurons) and small enough so that the region considered is homogeneous.
Our group is developing mathematical and numerical methods allowing on one hand to produce dynamic mean-field equations from the physiological characteristics of neural structure (neurons type, synapse type and anatomical connectivity between neurons populations), and on the other so simulate these equations. These methods use tools from advanced probability theory such as the theory of Large Deviations [7] and the study of interacting diffusions [1]. Our investigations have shown that the rigorous dynamics mean-field equations can have a quite more complex structure than the ones commonly used in the literature (e.g. [67]) as soon as realistic effects such as synaptic variability are taken into account. Our goal is to relate those theoretical results with experimental measurement, especially in the field of optical imaging. For this we are collaborating with Institut des Neurosciences de la Timone, Marseille.
Selected publications on this topic.

3.3. Neural fields

Neural fields are a phenomenological way of describing the activity of population of neurons by delay integro-differential equations. This continuous approximation turns out to be very useful to model large brain areas such as those involved in visual perception. The mathematical properties of these equations and their solutions are still imperfectly known, in particular in the presence of delays, different time scales and of noise.
Our group is developing mathematical and numerical methods for analysing these equations. These methods are based upon techniques from mathematical functional analysis [6], bifurcation theory [11], equivariant bifurcation analysis, delay equations, and stochastic partial differential equations. We have been able to characterize the solutions of these neural fields equations and their bifurcations, apply and expand the theory to account for such perceptual phenomena as edge, texture [3], and motion perception. We have also developed a theory of the delayed neural fields equations, in particular in the case of constant delays and propagation delays that must be taken into account when attempting to model large size cortical areas [82]. This theory is based on center manifold and normal forms ideas. We are currently extending the theory to take into account various sources of noise using tools from the theory of stochastic partial differential equations.

Selected publications on this topic.

3.4. Spike train statistics

The neuronal activity is manifested by the emission of action potentials (“spikes”) constituting spike trains. Those spike trains are usually not exactly reproducible when repeating the same experiment, even with a very good control ensuring that experimental conditions have not changed. Therefore, researchers are seeking models for spike train statistics, assumed to be characterized by a canonical probabilities giving the statistics of spatio-temporal spike patterns. A current goal in experimental analysis of spike trains is to approximate this probability from data. Several approach exist either based on (i) generic principles (maximum likelihood, maximum entropy); (ii) phenomenological models (Linear-Non linear, Generalized Linear Model, mean-field); (iii) Analytical results on spike train statistics in Neural Network models.

Our group is working on those 3 aspects, on a fundamental and on a practical (numerical) level. On one hand, we have published analytical (and rigorous) results on statistics of spike trains in canonical neural network models (Integrate and Fire, conductance based with chemical and electric synapses) [2], [54], [45]. The main result is the characterization of spike train statistics by a Gibbs distribution whose potential can be explicitly computed using some approximations. Note that this result does not require an assumption of stationarity. We have also shown that the distributions considered in the cases (i), (ii), (iii) above are all Gibbs distributions [55]. On the other hand, we are proposing new algorithms for data processing [25]. We have developed a C++ software for spike train statistics based on Gibbs distributions analysis and freely available at https://enas.inria.fr/. We are using this software in collaboration with several biologist groups involved in the analysis of retina spike trains (Centro de Neurociencia Valparaiso; Institut de la vision, Paris; Faculty of Medical Sciences, Newcastle University, Institute for Adaptive and Neural Computation, University of Edinburgh.
Selected publications on this topic.

3.5. Synaptic Plasticity

Neural networks show amazing abilities to evolve and adapt, and to store and process information. These capabilities are mainly conditioned by plasticity mechanisms, and especially synaptic plasticity, inducing a mutual coupling between network structure and neuron dynamics. Synaptic plasticity occurs at many levels of organization and time scales in the nervous system (Bienenstock, Cooper, and Munroe, 1982). It is of course involved in memory and learning mechanisms, but it also alters excitability of brain areas and regulates behavioral states (e.g. transition between sleep and wakeful activity). Therefore, understanding the effects of synaptic plasticity on neurons dynamics is a crucial challenge.

Our group is developing mathematical and numerical methods to analyse this mutual interaction. On one hand, we have shown that plasticity mechanisms, Hebbian-like or STDP, have strong effects on neuron dynamics complexity, such as dynamics complexity reduction, and spike statistics (convergence to a specific Gibbs distribution via a variational principle), resulting in a response-adaptation of the network to learned stimuli [77], [78], [56]. We are also studying the conjugated effects of synaptic and intrinsic plasticity in collaboration with H. Berry (Inria Beagle) and B. Delord, J. Naudé, ISIR team, Paris. On the other hand, we have pursued a geometric approach in which we show how a Hopfield network represented by a neural field with modifiable recurrent connections undergoing slow Hebbian learning can extract the underlying geometry of an input space [63]. We have also pursued an approach based on the ideas developed in the theory of slow-fast systems (in this case a set of neural fields equations) in the presence of noise and applied temporal averaging methods to recurrent networks of noisy neurons undergoing a slow and unsupervised modification of their connectivity matrix called learning [64].

Selected publications on this topic.

3.6. Visual neuroscience

Our group focuses on the visual system to understand how information is encoded and processed resulting in visual percepts. To do so, we propose functional models of the visual system using a variety of mathematical formalisms, depending on the scale at which models are built, such as spiking neural networks or neural fields. So far, our efforts have been focused on the study of retinal processing, edge and texture perception, motion integration at the level of V1 and MT cortical areas.

At the retina level, we are modeling its circuitry [14] and we are studying the statistics of the spike train output (see, e.g., the software ENAS https://enas.inria.fr/). Real cell recordings are also analysed in collaboration with Institut de la vision, ParisCentro de Neurociencia Valparaiso; Institut de la vision, Paris; Faculty of Medical Sciences, Newcastle University. For visual edges perception, we have used the theory of neural fields [13]. For visual textures perception, we have used a combination of neural fields theory and equivariant bifurcations theory [3]. At the level of V1-MT cortical areas, we have been investigating the temporal dynamics of motion integration for a wide range of visual stimuli [76], [79], [52], [9]. This work is done in collaboration with Institut des Neurosciences de la Timone, Marseille.

Selected publications on this topic.

3.7. Neuromorphic vision

From the simplest vision architectures in insects to the extremely complex cortical hierarchy in primates, it is fascinating to observe how biology has found efficient solutions to solve vision problems. Pioneers in computer vision had this dream to build machines that could match and perhaps outperform human vision. This goal has not been reached, at least not on the scale that was originally planned, but the field of computer vision has met many other challenges from an unexpected variety of applications and fostered entirely new scientific and technological areas such as computer graphics and medical image analysis. However, modelling and emulating with computers biological vision largely remains an open challenge while there are still many outstanding issues in computer vision.

Our group is working on neuromorphic vision by proposing bio-inspired methods following our progress in visual neuroscience. Our goal is to bridge the gap between biological and computer vision, by applying our visual neuroscience models to challenging problems from computer vision such as optical flow estimation [80], coding/decoding approaches [71], [72] or classification [60], [61].

Selected publications on this topic.

NEUROSYS Team

3. Research Program

3.1. Main Objectives

The main challenge in computational neuroscience is the high complexity of neural systems. The brain is a complex system and exhibits a hierarchy of interacting subunits. On a specific hierarchical level, such subunits evolve on a certain temporal and spatial scale. The interactions of small units on a low hierarchical level build up larger units on a higher hierarchical level evolving on a slower time scale and larger spatial scale. By virtue of the different dynamics on each hierarchical level, until today the corresponding mathematical models and data analysis techniques on each level are still distinct. Only few analysis and modeling frameworks are known which link successfully at least two hierarchical levels.
Once having extracted models for different description levels, typically they are applied to obtain simulated activity which is supposed to reconstruct features in experimental data. Although this approach appears straight-forward, it implies various difficulties. Usually the models involve a large set of unknown parameters which determine the dynamical properties of the models. To optimally reconstruct experimental features, it is necessary to formulate an inverse problem to extract optimally such model parameters from the experimental data. Typically this is a rather difficult problem due to the low signal-to-noise ratio in experimental brain signals. Moreover, the identification of signal features to be reconstructed by the model is not obvious in most applications. Consequently an extended analysis of the experimental data is necessary to identify the interesting data features. It is important to combine such a data analysis step with the parameter extraction procedure to achieve optimal results. Such a procedure depends on the properties of the experimental data and hence has to be developped for each application separately.

3.2. Challenges

Eventually the implementation of the models and analysis techniques achieved promises to be able to construct novel data monitor. This construction involves additional challenges and stipulates the contact to realistic environments. By virtue of the specific applications of the research, the close contact to hospitals and medical enterprises shall be established in a longer term in order to (i) gain deeper insight into the specific application of the devices and (ii) build specific devices in accordance to the actual need. Collaborations with local and national hospitals and the pharmaceutical industry already exist.

3.3. Research Directions

- From the microscopic to the mesoscopic scale:
  One research direction focusses on the *relation of single neuron activity* on the microscopic scale *to the activity of neuronal populations*. To this end, the team investigates the stochastic dynamics of single neurons subject to external random inputs and involving random microscopic properties, such as random synaptic strengths and probability distributions of spatial locations of membrane ion channels. Such an approach yields a stochastic model of single neurons and allows the derivation of a stochastic neural population model.
  This bridge between the microscopic and mesoscopic scale may be performed via two pathways. The analytical and numerical treatment of the microscopic model may be called a *bottom-up approach*, since it leads to a population activity model based on microscopic activity. This approach allows to compare theoretical neural population activity to experimentally obtained population activity. The *top-down approach* aims at extracting signal features from experimental data gained from neural populations which give insight into the dynamics of neural populations and the underlying microscopic activity. The work on both approaches represents a well-balanced investigation of the neural system based on the systems properties.

- From the mesoscopic to the macroscopic scale:
  The other research direction aims to link neural population dynamics to macroscopic activity and behaviour or, more generally, to phenomenological features. This link is more indirect but a very powerful approach to understand the brain, e.g., in the context of medical applications. Since real neural systems, such as in mammals, exhibit an interconnected network of neural populations, the team studies analytically and numerically the network dynamics of neural populations to gain deeper insight into possible phenomena, such as traveling waves or enhancement and diminution of certain neural rhythms. Electroencephalography (EEG) is a wonderful brain imaging technique to study the overall brain activity in real time noninvasively. However it is necessary to develop robust techniques based on stable features by investigating the time and frequency domains of brain signals. Two types of information are typically used in EEG signals: (i) transient events such as evoked potentials, spindles and K-complexes and (ii) the power in specific frequency bands.

NUMED Project-Team

3. Research Program

3.1. Multiscale propagation phenomena in biology

3.1.1. Project team positioning

The originality of our work is the quantitative description of propagation phenomena for some models including several scales. We are able to compute the macroscopic speed of propagation and the distribution with respect to the microscopic variable at relevant locations (*e.g.* the edge and the back of the front) in a wide variety of models.

Multiscale modeling of propagation phenomena raises a lot of interest in several fields of application. This ranges from shock waves in kinetic equations (Boltzmann, BGK, etc...), bacterial chemotactic waves, selection-mutation models with spatial heterogeneities, age-structured models for epidemiology or subdiffusive processes.

Earlier works generally focused on numerical simulations, hydrodynamic limits to average over the microscopic variable, or specific models with only local features, not suitable for most of the relevant models. Our contribution enables to derive the relevant features of propagation analytically, and far from the hydrodynamic regime for a wide range of models including nonlocal interaction terms.

Our recent understanding is closely related to the analysis of large deviations in multiscale dispersion equations, for which we give important contributions too.

These advances are linked to the work of other Inria teams (BANG, DRACULA, BEAGLE), and collaborators in mathematics, physics and theoretical biology in France, Austria and UK.

3.1.2. Recent results

Vincent Calvez has focused on the modelling and analysis of propagation phenomena in structured populations. This includes chemotactic concentration waves, transport-reaction equations, coupling between ecological processes (reaction-diffusion) and evolutionary processes (selection of the fittest trait, adaptation), evolution of age structured poulations, and anomalous diffusion.

He has also continued his work on the optimal control of monotone linear dynamical systems, using the Hamilton-Jacobi framework, and the weak KAM theory.

Emeric Bouin has defended his PhD on December 2nd, 2014. He has accomplished his work under the supervision of Vincent Calvez and Emmanuel Grenier. He has studied propagation phenomena in multiscale models. He has focused on some specific behaviours arising from the multiscale nature of the problem, which are not described by classical reaction-diffusion models. For example, he has discovered unexpected acceleration behaviour in kinetic reaction-transport equations (Bouin, Calvez and Nadin, Arch. Ration. Mech. Anal. 2014).

Laetitia Giraldi was a post-doctoral fellow funded by the ANR grant MODPOL under the supervision of Vincent Calvez. She studied thoroughly a biomechanical model for the growth of plant or yeats cells. This new model couples standard equations for the displacement of the cell wall under internal pressure, and a reaction-diffusion equation set on the membrane accounting for the growth pattern has a function of the cell geometry. A rigorous linear stability analysis of the growing spherical shape, together with the development of a stable numerical scheme opens the way to future research in the coupling between growth and geometry.

Alvaro Mateos Gonzalez has started a PhD on September 2014 under the supervision of Vincent Calvez, and Hugues Berry (BEAGLE). He has already collaborated fruitfully with Thomas Lepoutre (DRACULA) and Hughes Berry to investigate the long-time asymptotics of a degenerate renewal equation. This is a first step towards the mathematical analysis of anomalous diffusion processes.

3.1.3. Collaborations

- Mathematical description of bacterial chemotactic waves:
  - **N. Bournaveas** (Univ. Edinburgh), **V. Calvez** (ENS de Lyon, Inria NUMED) **B. Perthame** (Univ. Paris 6, Inria BANG), **Ch. Schmeiser** (Univ. Vienna), **N. Vauchelet**: design of the model, analysis of traveling waves, analysis of optimal strategies for bacterial foraging.
  - **J. Saragosti**, **V. Calvez** (ENS de Lyon, Inria NUMED), **A. Buguin**, **P. Silberzan** (Institut Curie, Paris): experiments, design of the model, identification of parameters.
  - **F. Filbet**, **C. Yang** (Univ. Lyon 1): numerical simulations in 2D in curved geometries.
- Transport-reaction waves and large deviations:
  - **E. Bouin**, **V. Calvez** (ENS de Lyon, Inria NUMED), **E. Grenier** (ENS de Lyon, Inria NUMED), **G. Nadin** (Univ. Paris 6)
- Selection-mutation models of invasive species:
  - **E. Bouin** (ENS de Lyon, Inria NUMED), **V. Calvez** (ENS de Lyon, Inria NUMED), **S. Mirrahimi** (Inst. Math. Toulouse): construction of traveling waves, asymptotic propagation of fronts,
  - **E. Bouin** (ENS de Lyon, Inria NUMED), **V. Calvez** (ENS de Lyon, Inria NUMED), **N. Meunier**, (Univ. Paris 5), **B. Perthame** (Univ. Paris 6, Inria Bang), **G. Raoul** (CEFE, Montpellier), **R. Voituriez** (Univ. Paris 6): formal analysis, derivation of various asymptotic regimes.
- Age-structured equations for subdiffusive processes (just starting)
  - **H. Berry** (Inria BEAGLE), **V. Calvez** (ENS de Lyon, Inria NUMED), **Th. Lepoutre** (Inria DRACULA), **P. Gabriel** (Univ. UVSQ)

This work is also supportet by a PEPS project (CNRS) "Physique Théorique et ses Interfaces", led by N. Vauchelet (Univ. Paris 6).

3.2. Growth of biological tissues

3.2.1. Project-team positioning

The originality of our work is the derivation, analysis and numerical simulations of mathematical model for growing cells and tissues. This includes mechanical effects (growth induces a modification of the mechanical stresses) and biological effects (growth is potentially influenced by the mechanical forces).

This leads to innovative models, adapted to specific biological problems (*e.g.* suture formation, cell polarisation), but which share similar features. We perform linear stability analysis, and look for pattern formation issues (at least instability of the homogeneous state).

The biophysical literature of such models is large. We refer to the groups of Ben Amar (ENS Paris), Boudaoud (ENS de Lyon), Mahadevan (Harvard), etc.

Our team combines strong expertise in reaction-diffusion equations (V. Calvez) and mechanical models (P. Vigneaux). We develop linear stability analysis on evolving domains (due to growth) for coupled biomechanical systems.

Another direction of work is the mathematical analysis of classical tumor growth models. These continuous mechanics models are very close to classical equations like Euler or Navier Stokes equations in fluid mechanics. However they bring there own difficulties: Darcy law, multispecies equations, non newtonian dynamics (Bingham flows). Part of our work consist in deriving existence results and designing acute numerical schemes for these equations.

3.2.2. Recent results

We have worked on several biological issues. Cell polarisation is the main one. We first analyzed a nonlinear model proposed by theoretical physicists and biologists to describe spontaneous polarisation of the budding yeast *S. cerevisae*. The model assumes a dynamical transport of molecules in the cytoplasm. It is analogous to the Keller-Segel model for cell chemotaxis, except for the source of the transport flux. We developped nonlinear analysis and entropy methods to investigate pattern formation (Calvez et al 2012). We are currently validating the model on experimental data. The analysis of polarization of a single cell is a preliminary step before the study of mating in a population of yeast cells. In the mating phase, secretion of pheromones induces a dialogue between cells of opposite types.

We also derive realistic models for the growth of the fission yeast *S. pombe*. We proposed two models which couple growth and geometry of the cell. We aim to tackle the issue of pattern formation, and more specifically the instability of the spherical shape, leading to a rod shape. The mechanical coupling involves the distribution of microtubules in the cytoplasm, which bring material to the cell wall.

Over the evaluation period, Paul Vigneaux developped expertise in modelling and design of new numerical schemes for complex fluid models of the viscoplastic type. Associated materials are involved in a broad range of applications ranging from chemical industry to geophysical and biological materials. In the context of NUMED, this expertise is linked to the development of complex constitutive laws for cancer cell tissue. During the period, NUMED used mixed compressible/incompressible fluid model for tumor growth and viscoelastic fluid model. Viscoplastic is one of the other types of complex fluid model which is usable in the field. Mathematically, it involves variational inequalities and the need for specific numerical methods.

3.2.3. Collaborations

- **V. Calvez** (ENS de Lyon, Inria NUMED), **Th. Lepoutre** (Inria DRACULA), **N. Meunier**, (Univ. Paris 5), **N. Muller** (Univ. Paris 5), **P. Vigneaux** (ENS de Lyon, Inria NUMED): mathematical analysis of cell polarisation, numerical simulations
- **V. Calvez** (ENS de Lyon, Inria NUMED), **N. Meunier**, (Univ. Paris 5), **M. Piel**, (Institut Curie, Paris), **R. Voituriez** (Univ. Paris 6): biomechanical modeling of the growth of *S. pombe*
- **D. Bresch** (Univ. Chambéry), **V. Calvez** (ENS de Lyon, Inria NUMED), **R.H. Khonsari** (King's College London, CHU Nantes), **J. Olivier** (Univ. Aix-Marseille), **P. Vigneaux** (ENS de Lyon, Inria NUMED): modeling, analysis and simulations of suture formation.
- **Didier Bresch** (Univ Chambéry), **Benoit Desjardins**(Moma group): petrology.

ANR JCJC project "MODPOL", *Mathematical models for cell polarization*, led by Vincent Calvez (ENS de Lyon, CNRS, Inria NUMED).

3.3. Multiscale models in oncology

3.3.1. Project-team positioning

Since 15 years, the development of mathematical models in oncology has become a significant field of research throughout the world. Several groups of researchers in biomathematics have developed complex and multiscale continuous and discrete models to describe the pathological processes as well as the action of anticancer anti-cancer drugs. Many groups in US (e.g. Alexander Anderson's lab, Kristin Swansson's lab) and in Canada (e.g. Thomas Hillen, Gerda de Vries), quickly developed and published interesting modeling frameworks. The setup of European networks such as the Marie Curie research and training networks managed by Nicolas Bellomo and Luigi Preziosi constituted a solid and fertile ground for the development of new oncology models by teams of biomathematicians and in particular Zvia Agur (Israel), Philip Maini (UK), Helen Byrne (UK), Andreas Deutsch (Germany), or Miguel Herrero (Spain).

3.3.2. Results

We have worked on the development of a multiscale system for modeling the complexity of the cancer disease and generate new hypothesis on the use of anti-cancer drugs. This model relies on a multiscale formalism integrating a subcellular level integrating molecular interactions, a cell level (integrating the regulation of the cell cycle at the levels of individual cells) and a macroscopic level for describing the spatio-temporal dynamics of different types of tumor tissues (proliferating, hypoxic and necrotic). The model is thus composed by a set of partial differential equations (PDEs) integrating molecular network up to tissue dynamics using lax from fluid dynamic. This formalism is useful to investigate theoretically different cancer processes such as the angiogenesis and invasion. We have published several examples and case studies of the use of this model in particular, the action of phase-specific chemotherapies (Ribba, You et al. 2009), the use of anti-angiogenic drugs (Billy, Ribba et al. 2009) and their use in combination with chemotherapies (Lignet, Benzekry et al. 2013). This last work also integrates a model of the VEGF molecular pathway for proliferation and migration of endothelial cells in the context of cancer angiogenesis (Lignet, Calvez et al. 2013).

If these types of models present interesting framework to theoretically investigate biological hypothesis, they however present limitation due to their large number of parameters. In consequence, we decided to stop the development of the multiscale platform until exploration of alternative modeling strategies to deal with real data. We focus our interest on the use of mixed-effect modeling techniques as classically used in the field of pharmacokinetic and pharmacodynamics modeling. The general principal of this approach lies in the integration of several levels of variability in the model thus allowing for the simultaneous analysis of data in several individuals. Nowadays, complex algorithms allow for dealing with this problem when the model is composed by few ordinary differential equations (ODEs). However, no similar parameter estimation method is available for models defined as PDEs. In consequence, we decided: 1. To develop more simple models, based on systems of ODEs, assuming simplistic hypothesis of tumor growth and response to treatment but with a real focus on model ability to predict real data. 2. To work alone the development of parameter estimation methods for PDE models in oncology.

3.4. Parametrization of complex systems

3.4.1. Project-team positioning

We focus on a specific problem: the "population" parametrization of a complex system. More precisely, instead of trying to look for parameters in order to fit the available data for one patient, in many cases it is more pertinent to look for the distribution of the parameters (assuming that it is gaussian or log gaussian) in a population of patients, and to maximize the likelihood of the observations of all patients. It is a very useful strategy when few data per patients are available, but when we have a lot of patients. The number of parameters to find is multiplied by two (average and standard deviation for each parameter) but the number of data is greatly increased.

This strategy, that we will call "population" parametrization has been initiated in the eighties by software like Nonmem. Recently Marc Lavielle (Popix team) made a series of breakthroughs and designed a new powerfull algorithm, leading to Monolix software.

However population parametrization is very costly. It requires several hundred of thousands of model evaluations, which may be very long.

3.4.2. Results

We address the problem of computation time when the complex model is long to evaluate. In simple cases like reaction diffusion equations in one space dimension, the evaluation of the model may take a few seconds of even a few minutes. In more realistic geometries, the computation time would be even larger and can reach the hour or day. It is therefore impossible to run a SAEM algorith on such models, since it would be much too long. Moreover the underlying algorithm can not be parallelized.

We propose a new iterative approach combining a SAEM algorithm together with a kriging. This strategy appears to be very efficient, since we were able to parametrize a PDE model as fast as a simple ODE model.

We are currently developing the corresponding software.

3.5. Models for the analysis of efficacy data in oncology

3.5.1. Project-team positioning

The development of new drugs for oncology patients faces significant issues with a global attrition rate of 95 percents and only 40 percents of drug approval in phase III after successful phase II. As for meteorology, the analysis through modeling and simulation (MS), of time-course data related to anticancer drugs efficacy and/or toxicity constitutes a rational method for predicting drugs efficacy in patients. This approach, now supported by regulatory agencies such as the FDA, is expected to improve the drug development process and in consequence the treatment of cancer patients. A private company, Pharsight, has nowadays the leader team in the development of such modeling frameworks. In 2009, this team published a model describing tumor size time-course in more than one thousand colorectal cancer patients. This model was used in an MS framework to predict the outcome of a phase III clinical trials based on the analysis of phase II data. From 2009 to 2013, 12 published articles address similar analysis of different therapeutic indications such as lung, prostate, thyroid and renal cancer. A similar modeling activity is also proposed for the analysis of data in preclinical experiments, and in particular, experiments in mice. Animal experiments represent critical stages to decide if a drug molecule should be tested in humans. MS methods are considered as tools to better investigate the mechanisms of drug action and to potentially facilitate the transition towards the clinical phases of the drug development process. Our team has worked in the development of two modeling frameworks with application in both preclinical and clinical oncology. For the preclinical context, we have worked on the development of models focusing on the process of tumor angiogenesis, i.e. the formation of intra-tumoral blood vessels. At the clinical level, we have developed a model to predict tumor size dynamics in patients with low-grade glioma.

At Inria, several project-teams have developed similar efforts. The project-team BANG has a solid experience in the development of age-structured models of the cell cycle and tissue regulation of tumors with clinical applications for chronotherapy. BANG is also currently applying these types of partial differential equation (PDE) models to the study of leukemia through collaboration with the project-team DRACULA. Project-team MC2 has recently shown that the analysis, through a simplified PDE model of tumor growth and treatment response, of 3D imaging, could lead to correct prediction of tumor volume evolution in patients with pulmonary metastasis from thyroid cancer. Regarding specifically the modeling of brain tumors, project-team ASCLEPIOS has brought an important contribution towards personalized medicine in analyzing 3D data information from MRI with a multiscale model that describes the evolution of high grade gliomas in the brain. Their framework relies on the cancer physiopathological model that was mainly developed by Kristin Swanson and her group at the university of Washington.

Outside from Inria, we wish to mention here the work of the group of Florence Hubert in Marseille in the development of models with an interesting compromise between mathematical complexity and data availability. A national ANR project led by the team is expected to support the development of an MS methodology for the analysis of tumor size data in patients with metastases.

3.5.2. Results

Regarding our contribution in preclinical modeling, we have developed a model to analyze the dynamics of tumor progression in nude mice xenografted with HT29 or HCT116 colorectal cancer cells. This model, based on a system of ordinary differential equations (ODEs), integrated the different types of tumor tissues, and in particular, the proliferating, hypoxic and necrotic tissues. Practically, in our experiment, tumor size was periodically measured, and percentages of hypoxic and necrotic tissue were assessed using immunohistochemistry techniques on tumor samples after euthanasia. In the proposed model, the peripheral non-hypoxic tissue proliferates according to a generalized-logistic equation where the maximal tumor size is represented by a variable called "carrying capacity". The ratio of the whole tumor size to the carrying capacity was used to define the hypoxic stress. As this stress increases, non-hypoxic tissue turns hypoxic. Hypoxic tissue does not stop proliferating, but hypoxia constitutes a transient stage before the tissue becomes necrotic. As the tumor grows, the carrying capacity increases owing to the process of angiogenesis (Ribba, Watkin et al. 2011). The model

is shown to correctly predict tumor growth dynamics as well as percentages of necrotic and hypoxic tissues within the tumor.

Regarding our contribution in clinical oncology, we developed an ODE model based on the analysis of mean tumor diameter (MTD) time-course in low-grade glioma patients (Ribba, Kaloshi et al. 2012).

In this model, the tumor is composed of proliferative (P) and non-proliferative quiescent tissue (Q) expressed in millimeters. The proportion of proliferative tissue transitioning into quiescence is constant. The treatment directly eliminates proliferative cells by inducing lethal DNA damage while these cells progress through the cell cycle. The quiescent cells are also affected by the treatment and become damaged quiescent cells (k_{PQ}). Damaged quiescent cells, when re-entering the cell cycle, can repair their DNA and become proliferative once again (transition from Q_P to P) or can die due to unrepaired damages. We modeled the pharmacokinetics of the PCV chemotherapy using a kinetic-pharmacodynamic (K-PD) approach, in which drug concentration is assumed to decay according to an exponential function. In this model, we did not consider the three drugs separately. Rather, we assumed the treatment to be represented as a whole by a unique variable (C), which represents the concentration of a virtual drug encompassing the three chemotherapeutic components of the PCV regimen. We modeled the exact number of treatment cycles administered by setting the value of C to 1 (arbitrary unit) at the initiation of each cycle(T_{Treat}): $C(T = T_{Treat}) = 1$.

The resulting model is as follows:

$$\begin{aligned}
\frac{dC}{dt} &= -KDE \times C \\
\frac{dP}{dt} &= \lambda_P P \left(1 - \frac{P^{\bigstar}}{K}\right) + k_{Q_pP} Q_p - k_{PQ} P - \gamma \times C \times KDE \times P \\
\frac{dQ}{dt} &= k_{PQ} P - \gamma \times C \times KDE \times Q \\
\frac{dQ_p}{dt} &= \gamma \times C \times KDE \times Q - k_{Q_pP} Q_p - \delta_{Q_p} Q_p
\end{aligned} \tag{82}$$

We challenged this model with additional patient data. In particular, MTD time-course information from 24 patients treated with TMZ (subset of the 120 patients from SH) and 25 patients treated with radiotherapy (SH). Note that exactly the same K-PD approach was used to model treatment pharmacokinetic (including for radiotherapy). This choice, though not really realistic was adopted for simplicity reasons: the same model can be indifferently applied to the three different treatment modalities of LGG patients.

3.5.3. Collaborations

François Ducray and Jérôme Honnorat (Pierre Wertheimer Hospital in Lyon)

External support: grant INSERM PhysiCancer 2012 and Inria IPL MONICA

3.6. Stroke

3.6.1. Project team positioning

Stroke is a major public health problem since it represents the second leading cause of death worldwide and the first cause of acquired disability in adults.

Numed is currently starting completely new issues with D. Rousseau (INSA) and his team. We have now at hand a large data base of clinical images. Our aim is to develop model which are able to predict the final size of the dead brain area as a function of the first two clinical data.

PARIETAL Project-Team

3. Research Program

3.1. Inverse problems in Neuroimaging

Many problems in neuroimaging can be framed as forward and inverse problems. For instance, the neuroimaging *inverse problem* consists in predicting individual information (behavior, phenotype) from neuroimaging data, while the *forward problem* consists in fitting neuroimaging data with high-dimensional (e.g. genetic) variables. Solving these problems entails the definition of two terms: a loss that quantifies the goodness of fit of the solution (does the model explain the data reasonably well ?), and a regularization schemes that represents a prior on the expected solution of the problem. In particular some priors enforce some properties of the solutions, such as sparsity, smoothness or being piece-wise constant.
Let us detail the model used in the inverse problem: Let $\mathbf{X}$ be a neuroimaging dataset as an (n_{subj}, n_{voxels}) matrix, where n_{subj} and n_{voxels} are the number of subjects under study, and the image size respectively, $\mathbf{Y}$ an array of values that represent characteristics of interest in the observed population, written as (n_{subj}, n_f) matrix, where n_f is the number of characteristics that are tested, and β an array of shape (n_{voxels}, n_f) that represents a set of pattern-specific maps. In the first place, we may consider the columns $\mathbf{Y}_1, .., \mathbf{Y}_{n_f}$ of Y independently, yielding n_f problems to be solved in parallel:

$$\mathbf{Y}_i = \mathbf{X}\beta_i + \epsilon_i, \forall i \in \{1, .., n_f\},$$

where the vector contains β_i is the i^{th} row of β. As the problem is clearly ill-posed, it is naturally handled in a regularized regression framework:

$$\widehat{\beta_i} = \text{argmin}_{\beta_i} \|\mathbf{Y}_i - \mathbf{X}\beta_i\|^2 + \Psi(\beta_i), \tag{83}$$

where Ψ is an adequate penalization used to regularize the solution:

$$\Psi(\beta; \lambda_1, \lambda_2, \eta_1, \eta_2) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2 + \eta_1 \|\nabla\beta\|_1 + \eta_2 \|\nabla\beta\|_2 \tag{84}$$

with λ_1, λ_2, η_1, $\eta_2 \geq 0$ (this formulation particularly highlights the fact that convex regularizers are norms or quasi-norms). In general, only one or two of these constraints is considered (hence is enforced with a non-zero coefficient):

- When $\lambda_1 > 0$ only (LASSO), and to some extent, when $\lambda_1, \lambda_2 > 0$ only (elastic net), the optimal solution β is (possibly very) sparse, but may not exhibit a proper image structure; it does not fit well with the intuitive concept of a brain map.
- Total Variation regularization (see Fig. 1) is obtained for ($\eta_1 > 0$ only), and typically yields a piece-wise constant solution. It can be associated with Lasso to enforce both sparsity and sparse variations.
- Smooth lasso is obtained with ($\eta_2 > 0$ and $\lambda_1 > 0$ only), and yields smooth, compactly supported spatial basis functions.

The performance of the predictive model can simply be evaluated as the amount of variance in $\mathbf{Y}_i$ fitted by the model, for each $i \in \{1, .., n_f\}$. This can be computed through cross-validation, by *learning* $\widehat{\beta_i}$ on some part of the dataset, and then estimating $(Y_i - X\widehat{\beta_i})$ using the remainder of the dataset.

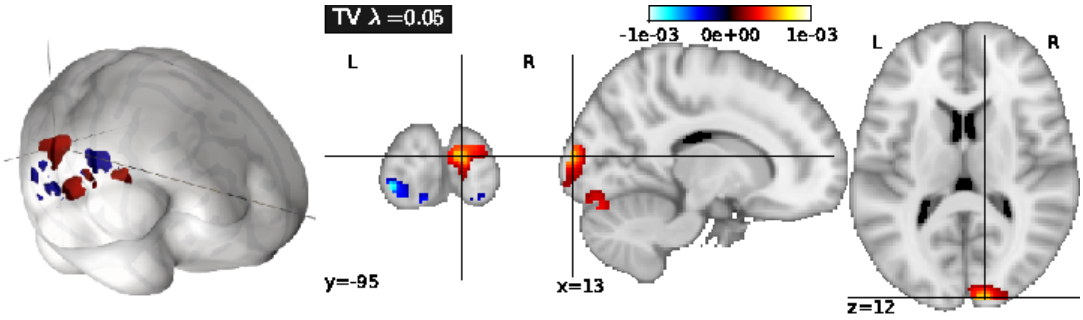


Figure 1. Example of the regularization of a brain map with total variation in an inverse problem. The problem here consists in predicting the spatial scale of an object presented as a stimulus, given functional neuroimaging data acquired during the observation of an image. Learning and test are performed across individuals. Unlike other approaches, Total Variation regularization yields a sparse and well-localized solution that enjoys particularly high accuracy.

This framework is easily extended by considering

- *Grouped penalization*, where the penalization explicitly includes a prior clustering of the features, i.e. voxel-related signals, into given groups. This is particularly important to include external anatomical priors on the relevant solution.
- *Combined penalizations*, i.e. a mixture of simple and group-wise penalizations, that allow some variability to fit the data in different populations of subjects, while keeping some common constraints.
- *Logistic regression*, where a logistic non-linearity is applied to the linear model so that it yields a probability of classification in a binary classification problem.
- *Robustness to between-subject variability* is an important question, as it makes little sense that a learned model depends dramatically on the particular observations used for learning. This is an important issue, as this kind of robustness is somewhat opposite to sparsity requirements.
- *Multi-task learning*: if several target variables are thought to be related, it might be useful to constrain the estimated parameter vector β to have a shared support across all these variables.
  For instance, when one of the variables $\mathbf{Y}_i$ is not well fitted by the model, the estimation of other variables $\mathbf{Y}_j, j \neq i$ may provide constraints on the support of β_i and thus, improve the prediction of $\mathbf{Y}_i$. Yet this does not impose constraints on the non-zero parameters of the parameters β_i.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{85}$$

then

$$\widehat{\beta} = \text{argmin}_{\beta=(\beta_i),i=1..n_f} \sum_{i=1}^{n_f} \|\mathbf{Y_i} - \mathbf{X}\beta_\mathbf{i}\|^2 + \lambda \sum_{j=1}^{n_{voxels}} \sqrt{\sum_{i=1}^{n_f} \beta_{\mathbf{i,j}}^{\mathbf{2}}} \tag{86}$$

3.2. Multivariate decompositions

Multivariate decompositions are an important tool to model complex data such as brain activation images: for instance, one might be interested in extracting an *atlas of brain regions* from a given dataset, such as regions depicting similar activities during a protocol, across multiple protocols, or even in the absence of protocol (during resting-state). These data can often be factorized into spatial-temporal components, and thus can be estimated through *regularized Principal Components Analysis* (PCA) algorithms, which share some common steps with regularized regression.
Let $\mathbf{X}$ be a neuroimaging dataset written as an (n_{subj}, n_{voxels}) matrix, after proper centering; the model reads

$$\mathbf{X} = \mathbf{AD} + \epsilon, \tag{87}$$

where $\mathbf{D}$ represents a set of n_{comp} spatial maps, hence a matrix of shape (n_{comp}, n_{voxels}), and $\mathbf{A}$ the associated subject-wise loadings. While traditional PCA and independent components analysis are limited to reconstruct components $\mathbf{D}$ within the space spanned by the column of $\mathbf{X}$, it seems desirable to add some constraints on the rows of $\mathbf{D}$, that represent spatial maps, such as sparsity, and/or smoothness, as it makes the interpretation of these maps clearer in the context of neuroimaging.
This yields the following estimation problem:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{AD}\|^2 + \Psi(\mathbf{D}) \text{ s.t. } \|\mathbf{A}_i\| = 1 \; \forall i \in \{1..n_f\}, \tag{88}$$

where $(\mathbf{A}_i),\ i \in \{1..n_f\}$ represents the columns of $\mathbf{A}$. Ψ can be chosen such as in Eq. (2) in order to enforce smoothness and/or sparsity constraints.
The problem is not jointly convex in all the variables but each penalization given in Eq (2) yields a convex problem on $\mathbf{D}$ for $\mathbf{A}$ fixed, and conversely. This readily suggests an alternate optimization scheme, where $\mathbf{D}$ and $\mathbf{A}$ are estimated in turn, until convergence to a local optimum of the criterion. As in PCA, the extracted components can be ranked according to the amount of fitted variance. Importantly, also, estimated PCA models can be interpreted as a probabilistic model of the data, assuming a high-dimensional Gaussian distribution (probabilistic PCA).

3.3. Covariance estimation

Another important estimation problem stems from the general issue of learning the relationship between sets of variables, in particular their covariance. Covariance learning is essential to model the dependence of these variables when they are used in a multivariate model, for instance to assess whether an observation is aberrant or not or in classification problems. Covariance learning is necessary to model latent interactions in high-dimensional observation spaces, e.g. when considering multiple contrasts or functional connectivity data.
The difficulties are two-fold: on the one hand, there is a shortage of data to learn a good covariance model from an individual subject, and on the other hand, subject-to-subject variability poses a serious challenge to the use of multi-subject data. While the covariance structure may vary from population to population, or depending on the input data (activation versus spontaneous activity), assuming some shared structure across problems, such as their sparsity pattern, is important in order to obtain correct estimates from noisy data. Some of the most important models are:

- **Sparse Gaussian graphical models**, as they express meaningful conditional independence relationships between regions, and do improve conditioning/avoid overfit.
- **Decomposable models**, as they enjoy good computational properties and enable intuitive interpretations of the network structure. Whether they can faithfully or not represent brain networks is an important question that needs to be addressed.
- **PCA-based regularization of covariance** which is powerful when modes of variation are more important than conditional independence relationships.

Adequate model selection procedures are necessary to achieve the right level of sparsity or regularization in covariance estimation; the natural evaluation metric here is the out-of-samples likelihood of the associated Gaussian model. Another essential remaining issue is to develop an adequate statistical framework to test differences between covariance models in different populations. To do so, we consider different means of parametrizing covariance distributions and how these parametrizations impact the test of statistical differences across individuals.

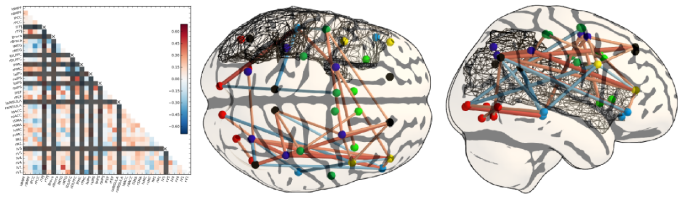


Figure 2. Example of functional connectivity analysis: The correlation matrix describing brain functional connectivity in a post-stroke patient (lesion volume outlined as a mesh) is compared to a group of control subjects. Some edges of the graphical model show a significant difference, but the statistical detection of the difference requires a sophisticated statistical framework for the comparison of graphical models.

POMDAPI Project-Team (section vide)

POPIX Team

3. Research Program

3.1. Research Program

Mathematical models that characterize complex biological phenomena are complex numerical models which are defined by systems of ordinary differential equations when dealing with dynamical systems that evolve with respect to time, or by partial differential equations when there is a spatial component to the model. Also, it is sometimes useful to integrate a stochastic aspect into the dynamical systems in order to model stochastic intra-individual variability.

In order to use such methods, we are rapidly confronted with complex numerical difficulties, generally related to resolving the systems of differential equations. Furthermore, to be able to check the quality of a model, we require data. The statistical aspect of the model is thus critical in its way of taking into account different sources of variability and uncertainty, especially when data comes from several individuals and we are interested in characterizing the inter-subject variability. Here, the tool of reference is mixed-effects models.

Mixed-effects models are statistical models with both fixed effects and random effects, i.e., mixed effects. They are useful in many real-world situations, especially in the physical, biological and social sciences. In particular, they are well-adapted to situations where repeated measurements are made on the same individual/statistical unit.

POPIX develops new methods for estimation of complex mixed-effects models. Some of the extensions to these models that POPIX is actively researching include:

- models defined by a large system of differential equations
- models defined by a system of stochastic differential equations
- models defined by partial differential equations
- mixed hidden Markov models
- mixture models and model mixtures
- time-to-event models
- models including a large number of covariates

It is also important to clarify that POPIX is not meant to be a team of modelers; our main activity is not to develop models, but to develop tools for modelers. Indeed, we are of course led via our various collaborations to interact closely with modelers involved in model development, in particular in the case of our collaborations with modeling and simulation teams in the pharmaceutical industry. But POPIX is not in the business of building PKPD models per se.

Lastly, though pharmacometrics remains the main field of interest for the population approach, this approach is also appropriate to address other types of complex biological phenomena exhibiting inter-individual variability and necessitating therefore to be described by numerical and statistical models. We have already demonstrated the relevance of the developed approaches and tools in diverse other domains such as agronomy for characterizing corn production, and cellular biology for characterizing the cell cycle and the creation of free radicals in cells. Now we wish to push on to explore new areas of modeling such as for the respiratory system and blood flow. But again, it is not within the scope of the activities of POPIX to develop new models; instead, the goal is to demonstrate the relevance of the population approach in these areas.

REO Project-Team

3. Research Program

3.1. Multiphysics modeling

In large vessels and in large bronchi, blood and air flows are generally supposed to be governed by the incompressible Navier-Stokes equations. Indeed in large arteries, blood can be supposed to be Newtonian, and at rest air can be modeled as an incompressible fluid. The cornerstone of the simulations is therefore a Navier-Stokes solver. But other physical features have also to be taken into account in simulations of biological flows, in particular fluid-structure interaction in large vessels and transport of sprays, particles or chemical species.

3.1.1. Fluid-structure interaction

Fluid-structure coupling occurs both in the respiratory and in the circulatory systems. We focus mainly on blood flows since our work is more advanced in this field. But the methods developed for blood flows could be also applied to the respiratory system.

Here "fluid-structure interaction" means a coupling between the 3D Navier-Stokes equations and a 3D (possibly thin) structure in large displacements.

The numerical simulations of the interaction between the artery wall and the blood flows raise many issues: (1) the displacement of the wall cannot be supposed to be infinitesimal, geometrical nonlinearities are therefore present in the structure and the fluid problem have to be solved on a moving domain (2) the densities of the artery walls and the blood being close, the coupling is strong and has to be tackled very carefully to avoid numerical instabilities, (3) "naive" boundary conditions on the artificial boundaries induce spurious reflection phenomena.

Simulation of valves, either at the outflow of the cardiac chambers or in veins, is another example of difficult fluid-structure problems arising in blood flows. In addition, very large displacements and changes of topology (contact problems) have to be handled in those cases.

Due to stability reasons, it seems impossible to successfully apply in hemodynamics the explicit coupling schemes used in other fluid-structure problems, like aeroelasticity. As a result, fluid-structure interaction in biological flows raise new challenging issues in scientific computing and numerical analysis : new schemes have to be developed and analyzed.

We have proposed and analyzed over the last few years several efficient fluid-structure interaction algorithms. This topic remains very active. We are now using these algorithms to address inverse problems in blood flows to make patient specific simulations (for example, estimation of artery wall stiffness from medical imaging).

3.1.2. Aerosol

Complex two-phase fluids can be modeled in many different ways. Eulerian models describe both phases by physical quantities such as the density, velocity or energy of each phase. In the mixed fluid-kinetic models, the biphasic fluid has one dispersed phase, which is constituted by a spray of droplets, with a possibly variable size, and a continuous classical fluid.

This type of model was first introduced by Williams [68] in the frame of combustion. It was later used to develop the Kiva code [58] at the Los Alamos National Laboratory, or the Hesione code [63], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [61], therapeutic sprays, *etc.* One of the interests of such a model is that various phenomena on the droplets can be taken into account with an accurate precision: collision, breakups, coagulation, vaporization, chemical reactions, *etc.*, at the level of the droplets.

The model usually consists in coupling a kinetic equation, that describes the spray through a probability density function, and classical fluid equations (typically Navier-Stokes). The numerical solution of this system relies on the coupling of a method for the fluid equations (for instance, a finite volume method) with a method fitted to the spray (particle method, Monte Carlo).

We are mainly interested in modeling therapeutic sprays either for local or general treatments. The study of the underlying kinetic equations should lead us to a global model of the ambient fluid and the droplets, with some mathematical significance. Well-chosen numerical methods can give some tracks on the solutions behavior and help to fit the physical parameters which appear in the models.

3.2. Multiscale modeling

Multiscale modeling is a necessary step for blood and respiratory flows. In this section, we focus on blood flows. Nevertheless, similar investigations are currently carried out on respiratory flows.

3.2.1. Arterial tree modeling

Problems arising in the numerical modeling of the human cardiovascular system often require an accurate description of the flow in a specific sensible subregion (carotid bifurcation, stented artery, *etc.*). The description of such local phenomena is better addressed by means of three-dimensional (3D) simulations, based on the numerical approximation of the incompressible Navier-Stokes equations, possibly accounting for compliant (moving) boundaries. These simulations require the specification of boundary data on artificial boundaries that have to be introduced to delimit the vascular district under study. The definition of such boundary conditions is critical and, in fact, influenced by the global systemic dynamics. Whenever the boundary data is not available from accurate measurements, a proper boundary condition requires a mathematical description of the action of the reminder of the circulatory system on the local district. From the computational point of view, it is not affordable to describe the whole circulatory system keeping the same level of detail. Therefore, this mathematical description relies on simpler models, leading to the concept of *geometrical multiscale* modeling of the circulation [64]. The underlying idea consists in coupling different models (3D, 1D or 0D) with a decreasing level of accuracy, which is compensated by their decreasing level of computational complexity.

The research on this topic aims at providing a correct methodology and a mathematical and numerical framework for the simulation of blood flow in the whole cardiovascular system by means of a geometric multiscale approach. In particular, one of the main issues will be the definition of stable coupling strategies between 3D and reduced order models.

To model the arterial tree, a standard way consists of imposing a pressure or a flow rate at the inlet of the aorta, *i.e.* at the network entry. This strategy does not allow to describe important features as the overload in the heart caused by backward traveling waves. Indeed imposing a boundary condition at the beginning of the aorta artificially disturbs physiological pressure waves going from the arterial tree to the heart. The only way to catch this physiological behavior is to couple the arteries with a model of heart, or at least a model of left ventricle.

A constitutive law for the myocardium, controlled by an electrical command, has been developed in the CardioSense3D project [0]. One of our objectives is to couple artery models with this heart model.

A long term goal is to achieve 3D simulations of a system including heart and arteries. One of the difficulties of this very challenging task is to model the cardiac valves. To this purpose, we investigate a mix of arbitrary Lagrangian Eulerian and fictitious domain approaches or x-fem strategies, or simplified valve models based on an immersed surface strategy.

[0] http://www-sop.inria.fr/CardioSense3D/

3.2.2. Heart perfusion modeling

The heart is the organ that regulates, through its periodical contraction, the distribution of oxygenated blood in human vessels in order to nourish the different parts of the body. The heart needs its own supply of blood to work. The coronary arteries are the vessels that accomplish this task. The phenomenon by which blood reaches myocardial heart tissue starting from the blood vessels is called in medicine perfusion. The analysis of heart perfusion is an interesting and challenging problem. Our aim is to perform a three-dimensional dynamical numerical simulation of perfusion in the beating heart, in order to better understand the phenomena linked to perfusion. In particular the role of the ventricle contraction on the perfusion of the heart is investigated as well as the influence of blood on the solid mechanics of the ventricle. Heart perfusion in fact implies the interaction between heart muscle and blood vessels, in a sponge-like material that contracts at every heartbeat via the myocardium fibers.

Despite recent advances on the anatomical description and measurements of the coronary tree and on the corresponding physiological, physical and numerical modeling aspects, the complete modeling and simulation of blood flows inside the large and the many small vessels feeding the heart is still out of reach. Therefore, in order to model blood perfusion in the cardiac tissue, we must limit the description of the detailed flows at a given space scale, and simplify the modeling of the smaller scale flows by aggregating these phenomena into macroscopic quantities, by some kind of "homogenization" procedure. To that purpose, the modeling of the fluid-solid coupling within the framework of porous media appears appropriate.

Poromechanics is a simplified mixture theory where a complex fluid-structure interaction problem is replaced by a superposition of both components, each of them representing a fraction of the complete material at every point. It originally emerged in soils mechanics with the work of Terzaghi [67], and Biot [59] later gave a description of the mechanical behavior of a porous medium using an elastic formulation for the solid matrix, and Darcy's law for the fluid flow through the matrix. Finite strain poroelastic models have been proposed (see references in [60]), albeit with *ad hoc* formulations for which compatibility with thermodynamics laws and incompressibility conditions is not established.

3.2.3. Tumor and vascularization

The same way the myocardium needs to be perfused for the heart to beat, when it has reached a certain size, tumor tissue needs to be perfused by enough blood to grow. It thus triggers the creation of new blood vessels (angiogenesis) to continue to grow. The interaction of tumor and its micro-environment is an active field of research. One of the challenges is that phenomena (tumor cell proliferation and death, blood vessel adaptation, nutrient transport and diffusion, etc) occur at different scales. A multi-scale approach is thus being developed to tackle this issue. The long term objective is to predict the efficiency of drugs and optimize therapy of cancer.

3.2.4. Respiratory tract modeling

We aim at developing a multiscale model of the respiratory tract. Intraprenchymal airways distal from generation 7 of the tracheabronchial tree (TBT), which cannot be visualized by common medical imaging techniques, are modeled either by a single simple model or by a model set according to their order in TBT. The single model is based on straight pipe fully developed flow (Poiseuille flow in steady regimes) with given alveolar pressure at the end of each compartment. It will provide boundary conditions at the bronchial ends of 3D TBT reconstructed from imaging data. The model set includes three serial models. The generation down to the pulmonary lobule will be modeled by reduced basis elements. The lobular airways will be represented by a fractal homogenization approach. The alveoli, which are the gas exchange loci between blood and inhaled air, inflating during inspiration and deflating during expiration, will be described by multiphysics homogenization.

SAGE Project-Team

3. Research Program

3.1. Numerical algorithms and high performance computing

Linear algebra is at the kernel of most scientific applications, in particular in physical or chemical engineering. For example, steady-state flow simulations in porous media are discretized in space and lead to a large sparse linear system. The target size is 10^7 in 2D and 10^{10} in 3D. For transient models such as diffusion, the objective is to solve about 10^4 linear systems for each simulation. Memory requirements are of the order of Giga-bytes in 2D and Tera-bytes in 3D. CPU times are of the order of several hours to several days. Several methods and solvers exist for large sparse linear systems. They can be divided into three classes: direct, iterative or semi-iterative. Direct methods are highly efficient but require a large memory space and a rapidly increasing computational time. Iterative methods of Krylov type require less memory but need a scalable preconditioner to remain competitive. Iterative methods of multigrid type are efficient and scalable, used by themselves or as preconditioners, with a linear complexity for elliptic or parabolic problems but they are not so efficient for hyperbolic problems. Semi-iterative methods such as subdomain methods are hybrid direct/iterative methods which can be good tradeoffs. The convergence of iterative and semi-iterative methods and the accuracy of the results depend on the condition number which can blow up at large scale. The objectives are to analyze the complexity of these different methods, to accelerate convergence of iterative methods, to measure and improve the efficiency on parallel architectures, to define criteria of choice.

In geophysics, a main concern is to solve inverse problems in order to fit the measured data with the model. Generally, this amounts to solve a linear or nonlinear least-squares problem. Complex models are in general coupled multi-physics models. For example, reactive transport couples advection-diffusion with chemistry. Here, the mathematical model is a set of nonlinear Partial Differential Algebraic Equations. At each timestep of an implicit scheme, a large nonlinear system of equations arise. The challenge is to solve efficiently and accurately these large nonlinear systems.

Approximation in Krylov subspace is in the core of the team activity since it provides efficient iterative solvers for linear systems and eigenvalue problems as well. The later are encountered in many fields and they include the singular value problem which is especially useful when solving ill posed inverse problems.

3.2. Numerical models applied to hydrogeology and physics

The team Sage is strongly involved in numerical models for hydrogeology and physics. There are many scientific challenges in the area of groundwater simulations. This interdisciplinary research is very fruitful with cross-fertilizing subjects. For example, high performance simulations were very helpful for finding out the asymptotic behaviour of the plume of solute transported by advection-dispersion. Numerical models are necessary to understand flow transfer in fractured media.

The team develops stochastic models for groundware simulations. Numerical models must then include Uncertainty Quantification methods, spatial and time discretization. Then, the discrete problems must be solved with efficient algorithms. The team develops parallel algorithms for complex numerical simulations and conducts performance analysis. Another challenge is to run multiparametric simulations. They can be multiple samples of a non intrusive Uncertainty Quantification method, or multiple samples of a stochastic method for inverse problems, or multiple samples for studying the sensitivity to a given model parameter. Thus these simulations are more or less independent and are well-suited to grid computing but each simulation requires powerful CPU and memory resources.

A strong commitment of the team is to develop the scientific software platform H2OLab for numerical simulations in heterogeneous hydrogeology.

SERPICO Project-Team

3. Research Program

3.1. Statistics and algorithms for computational microscopy

Many live-cell fluorescence imaging experiments are limited in time to prevent phototoxicity and photobleaching. The amount of light and time required to observe entire cell divisions can generate biological artifacts. In order to produce images compatible with the dynamic processes in living cells as seen in video-microscopy, we study the potential of denoising, superresolution, tracking, and motion analysis methods in the Bayesian and the robust statistics framework to extract information and to improve image resolution while preserving cell integrity.

In this area, we have already demonstrated that image denoising allows images to be taken more frequently or over a longer period of time [5]. The major advantage is to preserve cell integrity over time since spatio-temporal information can be restored using computational methods [6], [2], [7], [4]. This idea has been successfully applied to wide-field, spinning-disk confocal microscopy [1], TIRF [29], fast live imaging and 3D-PALM using the OMX system in collaboration with J. Sedat and M. Gustafsson at UCSF [5]. The corresponding ND-SAFIR denoiser software (see Section 5.1) has been licensed to a large set of laboratories over the world. New information restoration and image denoising methods are currently investigated to make SIM imaging compatible with the imaging of molecular dynamics in live cells. Unlike other optical sub-diffraction limited techniques (e.g. STED [43], PALM [31]) SIM has the strong advantage of versatility when considering the photo-physical properties of the fluorescent probes [40]. Such developments are also required to be compatible with "high-throughput microscopy" since several hundreds of cells are observed at the same time and the exposure times are typically reduced.

3.2. From image data to descriptors: dynamic analysis and trajectory computation

3.2.1. Motion analysis and tracking

The main challenge is to detect and track xFP tags with high precision in movies representing several Giga-Bytes of image data. The data are most often collected and processed automatically to generate information on partial or complete trajectories. Accordingly, we address both the methodological and computational issues involved in object detection and multiple objects tracking in order to better quantify motion in cell biology. Classical tracking methods have limitations as the number of objects and clutter increase. It is necessary to correctly associate measurements with tracked objects, i.e. to solve the difficult data association problem [52]. Data association even combined with sophisticated particle filtering techniques [55] or matching techniques [53] is problematic when tracking several hundreds of similar objects with variable velocities. Developing new optical flow and robust tracking methods and models in this area is then very stimulating since the problems we have to solve are really challenging and new for applied mathematics. In motion analysis, the goal is to formulate the problem of optical flow estimations in ways that take physical causes of brightness constancy violations into account [36], [41]. The interpretation of computed flow fields enables to provide spatio-temporal signatures of particular dynamic processes (e.g. Brownian and directed motion) and could help to complete the traffic modelling.

3.2.2. Event detection and motion classification

Protein complexes in living cells undergo multiple states of local concentration or dissociation, sometimes associated with diffusion processes. These events can be observed at the plasma membrane with TIRF microscopy. The difficulty arises when it becomes necessary to distinguish continuous motions due to trafficking from sudden events due to molecule concentrations or their dissociations. Typically, plasma membrane vesicle docking, membrane coat constitution or vesicle endocytosis are related to these issues.

Several approaches can be considered for the automatic detection of appearing and vanishing particles (or spots) in wide-field and TIRF microscopy images. Ideally this could be performed by tracking all the vesicles contained in the cell [55], [39]. Among the methods proposed to detect particles in microscopy images [57], [54], none is dedicated to the detection of a small number of particles appearing or disappearing suddenly between two time steps. Our way of handling small blob appearances/dis-appearances originates from the observation that two successive images are redundant and that occlusions correspond to blobs in one image which cannot be reconstructed from the other image [1] (see also [34]). Furthermore, recognizing dynamic protein behaviors in live cell fluorescence microscopy is of paramount importance to understand cell mechanisms. In our studies, it is challenging to classify intermingled dynamics of vesicular movements, docking/tethering, and ultimately, plasma membrane fusion of vesicles that leads to membrane diffusion or exocytosis of cargo proteins. Our aim is then to model, detect, estimate and classify subcellular dynamic events in TIRF microscopy image sequences. We investigate methods that exploits space-time information extracted from a couple of successive images to classify several types of motion (directed, diffusive (or Brownian) and confined motion) or compound motion.

3.3. From models to image data: simulation and modelling of membrane transport

Mathematical biology is a field in expansion, which has evolved into various branches and paradigms to address problems at various scales ranging from ecology to molecular structures. Nowadays, system biology [44], [59] aims at modelling systems as a whole in an integrative perspective instead of focusing on independent biophysical processes. One of the goals of these approaches is the cell in silico as investigated at Harvard Medical School (http://vcp.med.harvard.edu/) or the VCell of the University of Connecticut Health Center (http://www.nrcam.uchc.edu/). Previous simulation-based methods have been investigated to explain the spatial organization of microtubules [47] but the method is not integrative and a single scale is used to describe the visual patterns. In this line of work, we propose several contributions to combine imaging, traffic and membrane transport modelling in cell biology.

In this area, we focus on the analysis of transport intermediates (vesicles) that deliver cellular components to appropriate places within cells. We have already investigated the concept of Network Tomography (NT) [58] mainly developed for internet traffic estimation. The idea is to determine mean traffic intensities based on statistics accumulated over a period of time. The measurements are usually the number of vesicles detected at each destination region receiver. The NT concept has been investigated also for simulation [3] since it can be used to statistically mimic the contents of real traffic image sequences. In the future, we plan to incorporate more prior knowledge on dynamics to improve representation. An important challenge is to correlate stochastic, dynamical, one-dimensional *in silico* models studied at the nano-scale in biophysics, to 3D images acquired in vivo at the scale of few hundred nanometers. A difficulty is related to the scale change and statistical aggregation problems (in time and space) have to be handled.

SHACRA Project-Team

3. Research Program

3.1. Real-Time Biophysical Models

The principal objective of this scientific challenge is the modeling of the operative field, *i.e.* the anatomy and physiology of the patient that will be directly or indirectly targeted by a medical intervention. This requires to describe various phenomena such as soft-tissue deformation, fluid dynamics, electrical diffusion, or heat transfer. These models will help simulate the reaction of the patient's anatomy to the procedure, but also represent the behavior of complex organs such as the brain, the liver or the heart. A common requirement across these developments is the need for fast, possibly real-time, computation.

3.1.1. Real-time biomechanical modeling of solid structures

Soft tissue modeling holds a very important place in medical simulation. A large part of the realism of a simulation, in particular for surgery or laparoscopy simulation, relies upon the ability to describe soft tissue response during the simulated intervention. Several approaches have been proposed over the past ten years to model soft-tissue deformation in real-time (mainly for solid organs), usually based on elasticity theory and a finite element approach to solve the equations. We were among the first to propose an approach [3] using different computational strategies. Although significant improvements were obtained later on (for instance with the use of co-rotational methods to handle geometrical non-linearities) these works remain of limited clinical use as they essentially rely on linearized constitutive laws, and are rarely validated. An important part of our research remains dedicated to the development of new, more accurate models that are compatible with real-time computation. Such advanced models will not only permit to increase the realism of future training systems, but they will act as a bridge toward the development of patient-specific preoperative planning as well as augmented reality tools for the operating room.

3.1.2. Real-time biomechanical modeling of hollow structures

A large number of anatomical structures in the human body are vascularized (brain, liver, heart, kidneys, etc.) and recent interventions (such as interventional radiology procedures) rely on the vascular network as a therapeutical pathway. It is therefore essential to model the shape and deformable behavior of blood vessels. This can be done at two levels, depending of the objective. The global deformation of a vascular network can be represented using the vascular skeleton as a deformable (tree) structure, while local deformations need to be described using models of deformable surfaces. Other structures such as aneurysms, the colon or stomach can also benefit from being modeled as deformable surface, and we can rely on shell or thin plate theory to reach this objective.

3.1.3. Coupled physical models

Beyond biomechanical modeling of soft tissues, other physical phenomena have to be taken into account. In the context of percutaneous tumor ablation, both thermal and mechanical behaviors have to be modelled. Focusing especially on the simulation of cryoablation (freezing the pathological tissue), models for heat transfer have been implemented to simulate the evolution of temperature within living tissues. Pre-operative planning of the iceball can thus be envisaged. Moreover, this demonstrates the multi-physics aspect of SOFA.

3.1.4. Real-time electrophysiology

Electrophysiology plays an important role in the physiology of the human body, for instance by inducing muscles motion, and obviously through the nervous system. Also, many clinical procedures rely on electrical stimulation, such as defibrillation, neuromuscular or deep brain stimulation for instance. Yet, the modeling and the simulation of this phenomenon is still in its early stages. Our primary objective is to focus on cardiac electrophysiology, which plays a critical role in the understanding of heart mechanisms, and also in the planning of certain cardiac procedures. We propose to develop models and computational strategies aimed at real-time simulation, and to also provide personalization tools (parameter estimation) allowing to run patient-specific simulations from clinical data.

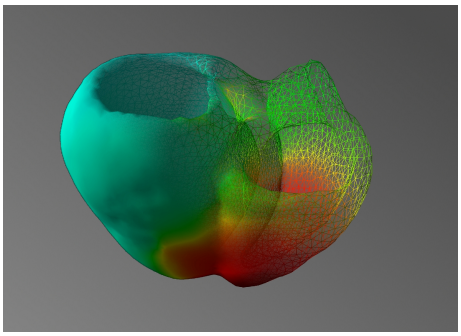


Figure 3. Patient-specific electrophysiology model of the human heart running in real-time

3.2. Numerical Methods for Complex Interactions

3.2.1. Dynamic topological changes

As mentioned previously, assisting the surgeon by providing either pre-operative planning or per-operative guidance assumes to increase the level of complexity and accuracy of our models, thus making the simulation more computationally-demanding. Innovative numerical methods must therefore be investigated. For instance, efforts are made to couple SOFA with CGoGN. CGoGN is a library based on combinatorial maps theory, specialized for representing and manipulating meshes. It is able to represent consistently objects of different dimensions composed of arbitrary cells (polygonal faces, polyhedral volumes). It provides an efficient way to explore the cells and their neighborhood; it allows to store data with the cells (both at execution time and compile time) and to efficiently modify the connectivity of the mesh even in highly dynamic cases. Adaptive meshing and efficient topological algorithm are thus available in SOFA.

3.2.2. Constraint models and boundary conditions

To simulate soft-tissue deformations accurately, the modeling technique must account for the intrinsic behavior of the modeled organ as well as for its biomechanical interactions with surrounding tissues or medical devices. While the biomechanical behavior of important organs (such as the brain or liver) has been studied extensively in the past, only few works exist dealing with the mechanical interactions between the anatomical structures. For tissue–tool interactions, most techniques rely on simple contact models, whereas advanced phenomena such as friction are rarely taken into account. While simplifications can produce plausible results in the case of interaction between the manipulator of a laparoscopic instrument and the surface of an organ, it is generally an insufficient approximation. As we move towards the simulations for planning or rehearsal, accurate modeling of contacts is playing an increasingly important role. For example, we have shown in [36] and [37] that complex interactions between a coil and an aneurysm, or alternatively between a flexible needle and a soft-tissue can be computed in real-time. In laparoscopic surgery, the main challenge is represented by modeling of interactions between anatomical structures rather than only between the instruments and the surface of the organ. Consequently, our objective was to model accurately the contacts with friction and other type on non-smooth interactions in a heterogeneous environment and to allow for stable haptic rendering. When different time integration strategies are used, another challenge is to compute the contact forces in such a way that integrity and stability of the overall simulation are maintained. Our objective was to propose a unified definition of such various boundary conditions and develop new numerical methods for simulations of heterogeneous objects.

3.3. Image-Driven Simulation: towards pre-operative planning and per-operative guidance

Image-guided therapy is a recent area of research that has the potential to bridge the gap between medical imaging and clinical routine by adapting pre-operative data to the time of the procedure. Several challenges are typically related to image-guided therapy, such as multi-modality image registration, which serves to align pre-operative images onto the patient. As most procedures deal with soft-tissues, elastic registration techniques are necessary to perform this step. Novel registration techniques began to account for soft tissue deformation using physically-based methods. Yet, several limitations still hinder the use of image-guided therapy in clinical routine. First, as registration methods become more complex, their computation time increases, thus lacking responsiveness. Second, as we have seen previously, many factors influence the deformation of soft-tissues, from patient-specific material properties to boundary conditions with surrounding anatomy. Another very similar, and related, problem is augmented reality, i.e. the real-time superposition of a virtual model onto the reality. In a clinical context, this can be very useful to help "see through" the anatomy. In this case, however, real-time registration of the virtual information onto the patient is mandatory. Our objective in this area is to combine our expertise in real-time soft-tissue modeling, complex interactions with image data to provide accurate and real-time registration, deformation, and tracking of virtual anatomical structures onto the patient.

The predictive capabilities of computer simulations may also be used to improve minimally invasive surgical procedures. While simulation results are sensitive to model parameters, initial and boundary conditions, we aim at combining computer-vision algorithms and simulation algorithms in order to produce dynamic data-driven simulation in clinical applications. The main idea is to use computer-vision algorithms from pre-operative diagnoses or per-operative video streams in order to extract meaningful data to feed the simulation engine and thus to increase the accuracy of the simulation. Clinical outcomes are expected in interventional radiology where the guidance is based on fluoroscopic imaging modality inducing high absorbed dose of X-rays for the patient and the clinical staff. In that context, using the prediction capabilities of the simulation may decrease the acquisition frequency of images, leading to a lower exposure of X-rays. Our objective in this area is to combine our expertise in patient-specific modeling and constraint models to achieve the dynamic coupling between images, pre-operative data and computer simulation.

SISTM Team

3. Research Program

3.1. Mecanistic modelling

When studying the dynamics of a given marker, say the HIV concentration in the blood (HIV viral load), one can for instance use descriptive models summarizing the dynamics over time in term of slopes of the trajectories [48]. These slopes can be compared between treatment groups or according to patients' characteristics. Another way for analyzing these data is to define a mathematical model based on the biological knowledge of what drives HIV dynamics. In this case, it is mainly the availability of target cells (the CD4+ T lymphocytes), the production and death rates of infected cells and the clearance of the viral particles that impact the dynamics. Then, a mathematical model most often based on ordinary differential equations (ODE) can be written [41]. Estimating the parameters of this model to fit observed HIV viral load gave a crucial insight in HIV pathogenesis as it revealed the very short half-life of the virions and infected cells and therefore a very high turnover of the virus, making mutations a very frequent event [40].

Having a good mechanistic model in a biomedical context such as HIV infection opens doors to various applications beyond a good understanding of the data. Global and individual predictions can be excellent because of the external validity of a model based on main biological mechanisms. Control theory may serve for defining optimal interventions or optimal designs to evaluate new interventions [33]. Finally, these models can capture explicitly the complex relationship between several processes that change over time and may therefore challenge other proposed approaches such as marginal structural models to deal with causal associations in epidemiology [32].

Therefore, we postulate that this type of model could be very useful in the context of our research that is in complex biological systems. The definition of the model needs to identify the parameter values that fit the data. In clinical research this is challenging because data are sparse, and often unbalanced, coming from populations of subjects. A substantial inter-individual variability is always present and needs to be accounted as this is the main source of information. Although many approaches have been developed to estimate the parameters of non-linear mixed models [45], [51], [36], [42], [37], [50], the difficulty associated with the complexity of ODE models and the sparsity of the data leading to identifiability issues need further research.

3.2. High dimensional data

With the availability of omics data such as genomics (DNA), transcriptomics (RNA) or proteomics (proteins), but also other types of data, such as those arising from the combination of large observational databases (e.g. in pharmacoepidemiology or environmental epidemiology), high-dimensional data have became increasingly common. Use of molecular biological technics such as Polymerase Chain Reaction (PCR) allows for amplification of DNA or RNA sequences. Nowadays, microarray and Next Generation Sequencing (NGS) techniques give the possibility to explore very large portions of the genome. Furthermore, other assays have also evolved, and traditional measures such as cytometry or imaging have became new sources of big data. Therefore, in the context of HIV research, the dimension of the datasets has much grown in term of number of variables per individual than in term of number of included patients although this latter is also growing thanks to the multi-cohort collaborations such as CASCADE or COHERE organized in the EuroCoord network [0]. As an exemple, in a recent phase 1/2 clinical trial evaluating the safety and the immunological response to a dendritic cell-based HIV vaccine, 19 infected patients were included. Bringing together data on cell count, cytokine production, gene expression and viral genome change led to a 20 Go database [19]. This is far from big databases faced in other areas but constitutes a revolution in clinical research where clinical trials of hundred of patients sized few hundred of Ko at most. Therefore, more than the storage and calculation capacities, the challenge is the comprehensive analysis of these datasets.

[0] see online at http://www.eurocoord.net

The objective is either to select the relevant information or to summarize it for understanding or prediction purposes. When dealing with high dimensional data, the methodological challenge arises from the fact that datasets typically contain many variables, much more than observations. Hence, multiple testing is an obvious issue that needs to be taken into account [46]. Furthermore, conventional methods, such as linear models, are inefficient and most of the time even inapplicable. Specific methods have been developed, often derived from the machine learning field, such as regularization methods [49]. The integrative analysis of large datasets is challenging. For instance, one may want to look at the correlation between two large scale matrices composed by the transcriptome in the one hand and the proteome on the other hand [38]. The comprehensive analysis of these large datasets concerning several levels from molecular pathways to clinical response of a population of patients needs specific approaches and a very close collaboration with the providers of data that is the immunologists, the virologists, the clinicians...

SISYPHE Project-Team (section vide)

STEEP Team

3. Research Program

3.1. Development of numerical systemic models (economy / society /environment) at local scales

The problem we consider is intrinsically interdisciplinary: it draws on social sciences, ecology or science of the planet. The modeling of the considered phenomena must take into account many factors of different nature which interact with varied functional relationships. These heterogeneous dynamics are *a priori* nonlinear and complex: they may have saturation mechanisms, threshold effects, and may be density dependent. The difficulties are compounded by the strong interconnections of the system (presence of important feedback loops) and multi-scale spatial interactions. Environmental and social phenomena are indeed constrained by the geometry of the area in which they occur. Climate and urbanization are typical examples. These spatial processes involve proximity relationships and neighborhoods, like for example, between two adjacent parcels of land, or between several macroscopic levels of a social organization. The multi-scale issues are due to the simultaneous consideration in the modeling of actors of different types and that operate at specific scales (spatial and temporal). For example, to properly address biodiversity issues, the scale at which we must consider the evolution of rurality is probably very different from the one at which we model the biological phenomena.

In this context, to develop flexible integrated systemic models (upgradable, modular, ...) which are efficient, realistic and easy to use (for developers, modelers and end users) is a challenge in itself. What mathematical representations and what computational tools to use? Nowadays many tools are used: for example, cellular automata (e.g. in the LEAM model), agent models (e.g. URBANSIM), system dynamics (e.g. World3), large systems of ordinary equations (e.g. equilibrium models such as TRANUS), and so on. Each of these tools has strengths and weaknesses. Is it necessary to invent other representations? What is the relevant level of modularity? How to get very modular models while keeping them very coherent and easy to calibrate? Is it preferable to use the same modeling tools for the whole system, or can we freely change the representation for each considered subsystem? How to easily and effectively manage different scales? (difficulty appearing in particular during the calibration process). How to get models which automatically adapt to the granularity of the data and which are always numerically stable? (this has also a direct link with the calibration processes and the propagation of uncertainties). How to develop models that can be calibrated with reasonable efforts, consistent with the (human and material) resources of the agencies and consulting firms that use them?

Before describing our research axes, we provide a brief overview of the types of models that we are or will be working with. As for LUTI (Land Use and Transportation Integrated) modeling, we have been using the TRANUS model since the start of our group. It is the most widely used LUTI model, has been developed since 1982 by the company Modelistica, and is distributed *via* Open Source software. TRANUS proceeds by solving a system of deterministic nonlinear equations and inequalities containing a number of economic parameters (e.g. demand elasticity parameters, location dispersion parameters, etc.). The solution of such a system represents an economic equilibrium between supply and demand. A second LUTI model that will be considered in the near future, within the CITiES project, is UrbanSim [0]. Whereas TRANUS aggregates over e.g. entire population or housing categories, UrbanSim takes a micro-simulation approach, modeling and simulating choices made at the level of individual households, businesses, and jobs, for instance, and it operates on a finer geographic scale than TRANUS.

[0] http://www.urbansim.org

On the other hand, the scientific domains related to eco-system services and ecological accounting are much less mature than the one of urban economy from a modelling point of view (as a consequence of our more limited knowledge of the relevant complex processes and/or more limited available data). Nowadays, the community working on ecological accounting and material flow analysis only proposes statistical models based on more or less simple data correlations. The eco-system service community has been using statical models too, but is also developing more sophisticated models based for example on system dynamics, multi-agent type simulations or cellular models. In the ESNET project, STEEP will work in particular on a land use/ land cover change (LUCC) modelling environments (LCM from Clark labs [0], and Dinamica [0]) which belongs to the category of spatially explicit statistical models.

In the following, our two main research axes are described, from the point of view of applied mathematical development. The domains of application of this research effort is described in the application section, where some details about the context of each field is given.

3.2. Model calibration and validation

The overall calibration of the parameters that drive the equations implemented in the above models is a vital step. Theoretically, as the implemented equations describe e.g. socio-economic phenomena, some of these parameters should in principle be accurately estimated from past data using econometrics and statistical methods like regressions or maximum likelihood estimates, e.g. for the parameters of logit models describing the residential choices of households. However, this theoretical consideration is often not efficient in practice for at least two main reasons. First, the above models consist of several interacting modules. Currently, these modules are typically calibrated independently; this is clearly sub-optimal as results will differ from those obtained after a global calibration of the interaction system, which is the actual final objective of a calibration procedure. Second, the lack of data is an inherent problem.

As a consequence, models are usually calibrated by hand. The calibration can typically take up to 6 months for a medium size LUTI model (about 100 geographic zones, about 10 sectors including economic sectors, population and employment categories). This clearly emphasizes the need to further investigate and at least semi-automate the calibration process. Yet, in all domains STEEP considers, very few studies have addressed this central issue, not to mention calibration under uncertainty which has largely been ignored (with the exception of a few uncertainty propagation analyses reported in the literature).

Besides uncertainty analysis, another main aspect of calibration is numerical optimization. The general state-of-the-art on optimization procedures is extremely large and mature, covering many different types of optimization problems, in terms of size (number of parameters and data) and type of cost function(s) and constraints. Depending on the characteristics of the considered models in terms of dimension, data availability and quality, deterministic or stochastic methods will be implemented. For the former, due to the presence of non-differentiability, it is likely, depending on their severity, that derivative free control methods will have to be preferred. For the latter, particle-based filtering techniques and/or metamodel-based optimization techniques (also called response surfaces or surrogate models) are good candidates.

These methods will be validated, by performing a series of tests to verify that the optimization algorithms are efficient in the sense that 1) they converge after an acceptable computing time, 2) they are robust and 3) that the algorithms do what they are actually meant to. For the latter, the procedure for this algorithmic validation phase will be to measure the quality of the results obtained after the calibration, i.e. we have to analyze if the calibrated model fits sufficiently well the data according to predetermined criteria.

To summarize, the overall goal of this research axis is to address two major issues related to calibration and validation of models: (a) defining a calibration methodology and developing relevant and efficient algorithms to facilitate the parameter estimation of considered models; (b) defining a validation methodology and developing the related algorithms (this is complemented by sensitivity analysis, see the following section). In both cases, analyzing the uncertainty that may arise either from the data or the underlying equations, and

[0] http://www.clarklabs.org/products/Land-Change-Modeler-Overview.cfm
[0] http://www.csr.ufmg.br/dinamica/

quantifying how these uncertainties propagate in the model, are of major importance. We will work on all those issues for the models of all the applied domains covered by STEEP.

3.3. Sensitivity analysis

A sensitivity analysis (SA) consists, in a nutshell, in studying how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model inputs. It is complementary to an uncertainty analysis, which focuses on quantifying uncertainty in model output. SA's can be useful for several purposes, such as guiding model development and identifying the most influential model parameters and critical data items. Identifying influential model parameters may help in divising metamodels (or, surrogate models) that approximate an original model and may be simulated, calibrated, or analyzed more efficiently. As for detecting critical data items, this may indicate for which type of data more effort must be spent in the data collection process in order to eventually improve the model's reliability. Finally, SA can be used as one means for validating models, together with validation based on historical data (or, put simply, using training and test data) and validation of model parameters and outputs by experts in the respective application area. All these uses of SA will be considered in our research.

The first two applications of SA are linked to model calibration, discussed in the previous section. Indeed, prior to the development of the calibration tools, one important step is to select the significant or sensitive parameters and to evaluate the robustness of the calibration results with respect to data noise (stability studies). This may be performed through a global sensitivity analysis, e.g. by computation of Sobol's indices. Many problems will have to be circumvented e.g. difficulties arising from dependencies of input variables, variables that obey a spatial organization, or switch inputs. We will take up on current work in the statistics community on SA for these difficult cases.

As for the third application of SA, model validation, a preliminary task bears on the propagation of uncertainties. Identifying the sources of uncertainties and their nature is crucial to propagate them via Monte Carlo techniques. To make a Monte Carlo approach computationally feasible, it is necessary to develop specific metamodels. Both the identification of the uncertainties and their propagation require a detailed knowledge of the data collection process; these are mandatory steps before a validation procedure based on SA can be implemented. First, we will focus on validating LUTI models, starting with the CITiES ANR project: here, an SA consists in defining various land use policies and transportation scenarios and in using these scenarios to test the integrated land use and transportation model. Current approaches for validation by SA consider several scenarios and propose various indicators to measure the simulated changes. We will work towards using sensitivity indices based on functional analysis of variance, which will allow us to compare the influence of various inputs on the indicators. For example it will allow the comparison of the influences of transportation and land use policies on several indicators.

TONUS Team

3. Research Program

3.1. Kinetic models for plasmas

The fundamental model for plasma physics is the coupled Vlasov-Maxwell kinetic model: the Vlasov equation describes the distribution function of particles (ions and electrons), while the Maxwell equations describe the electromagnetic field. In some applications, it may be necessary to take into account relativistic particles, which lead to consider the relativistic Vlasov equation, but generally, tokamak plasmas are supposed to be non relativistic. The particles distribution function depends on seven variables (three for space, three for velocity and one for time), which yields a huge amount of computations.

To these equations we must add several types of source terms and boundary conditions for representing the walls of the tokamak, the applied electromagnetic field that confines the plasma, fuel injection, collision effects, etc.

Tokamak plasmas possess particular features, which require developing specialized theoretical and numerical tools.

Because the magnetic field is strong, the particle trajectories have a very fast rotation around the magnetic field lines. A full resolution would require prohibitive amount of calculations. It is then necessary to develop models where the cyclotron frequency tends to infinity in order to obtain tractable calculations. The resulting model is called a gyrokinetic model. It allows us to reduce the dimensionality of the problem. Such models are implemented in GYSELA and Selalib. Those models require averaging of the acting fields during a rotation period along the trajectories of the particles. This averaging is called the gyroaverage and requires specific discretizations.

The tokamak and its magnetics fields present a very particular geometry. Some authors have proposed to return to the intrinsic geometrical versions of the Vlasov-Maxwell system in order to build better gyrokinetic models and adapted numerical schemes. This implies the use of sophisticated tools of differential geometry: differential forms, symplectic manifolds, and hamiltonian geometry.

In addition to theoretical modeling tools, it is necessary to develop numerical schemes adapted to kinetic and gyrokinetic models. Three kinds of methods are studied in TONUS: Particle-In-Cell (PIC) methods, semi-Lagrangian and fully Eulerian approaches.

3.1.1. Gyrokinetic models: theory and approximation

In most phenomena where oscillations are present, we can establish a three-model hierarchy: (i) the model parameterized by the oscillation period, (ii) the limit model and (iii) the Two-Scale model, possibly with its corrector. In a context where one wishes to simulate such a phenomenon where the oscillation period is small and where the oscillation amplitude is not small, it is important to have numerical methods based on an approximation of the Two-Scale model. If the oscillation period varies significantly over the domain of simulation, it is important to have numerical methods that approximate properly and effectively the model parameterized by the oscillation period and the Two-Scale model. Implemented Two-Scale Numerical Methods (for instance by Frénod et al. [36]) are based on the numerical approximation of the Two-Scale model. These are called of order 0. A Two-Scale Numerical Method is called of order 1 if it incorporates information from the corrector and from the equation to which this corrector is a solution. If the oscillation period varies between very small values and values of order 1, it is necessary to have new types of numerical schemes (Two-Scale Asymptotic Preserving Schemes of order 1 or TSAPS) with the property being able to preserve the asymptotics between the model parameterized by the oscillation period and the Two-Scale model with its corrector. A first work in this direction has been initiated by Crouseilles et al. [32].

3.1.2. Semi-Lagrangian schemes

The Strasbourg team has a long and recognized experience in numerical methods of Vlasov-type equations. We are specialized in both particle and phase space solvers for the Vlasov equation: Particle-in-Cell (PIC) methods and semi-Lagrangian methods. We also have a longstanding collaboration with the CEA of Cadarache for the development of the GYSELA software for gyrokinetic tokamak plasmas.

The Vlasov and the gyrokinetic models are partial differential equations that express the transport of the distribution function in the phase space. In the original Vlasov case, the phase space is the six-dimension position-velocity space. For the gyrokinetic model, the phase space is five-dimensional because we consider only the parallel velocity in the direction of the magnetic field and the gyrokinetic angular velocity instead of three velocity components.

A few years ago, Eric Sonnendrücker and his collaborators introduce a new family of methods for solving transport equations in the phase space. This family of methods are the semi-Lagrangian methods. The principle of these methods is to solve the equation on a grid of the phase space. The grid points are transported with the flow of the transport equation for a time step and interpolated back periodically onto the initial grid. The method is then a mix of particle Lagrangian methods and eulerian methods. The characteristics can be solved forward or backward in time leading to the Forward Semi-Lagrangian (FSL) or Backward Semi-Lagrangian (BSL) schemes. Conservative schemes based on this idea can be developed and are called Conservative Semi-Lagrangian (CSL).

GYSELA is a 5D full gyrokinetic code based on a classical backward semi-Lagrangian scheme (BSL) [43] for the simulation of core turbulence that has been developed at CEA Cadarache in collaboration with our team [37]. Although GYSELA was carefully developed to be conservative at lowest order, it is not exactly conservative, which might be an issue when the simulation is under-resolved, which always happens in turbulence simulations due to the formation of vortices which roll up.

3.1.3. PIC methods

Historically PIC methods have been very popular for solving the Vlasov equations. They allow solving the equations in the phase space at a relatively low cost. The main disadvantage of the method is that, due to its random aspect, it produces an important numerical noise that has to be controlled in some way, for instance by regularizations of the particles, or by divergence correction techniques in the Maxwell solver. We have a longstanding experience in PIC methods and we started implement them in SeLaLib. An important aspect is to adapt the method to new multicore computers. See the work by Crestetto and Helluy [31].

3.2. Reduced kinetic models for plasmas

As already said, kinetic plasmas computer simulations are very intensive, because of the gyrokinetic turbulence. In some situations, it is possible to make assumptions on the shape of the distribution function that simplify the model. We obtain in this way a family of fluid or reduced models.

Assuming that the distribution function has a Maxwellian shape, for instance, we obtain the MagnetoHydroDynamic (MHD) model. It is physically valid only in some parts of the tokamak (at the edges for instance). The fluid model is generally obtained from the hypothesis that the collisions between particles are strong. At Inria, fine collision models are mainly investigated in the KALIFFE team. In our approach we do not assume that the collisions are strong, but rather try to adapt the representation of the distribution function according to its shape, keeping the kinetic effects. The reduction is not necessarily a consequence of collisional effects. Indeed, even without collisions, the plasma may still relax to an equilibrium state over sufficiently long time scales (Landau damping effect). Recently, a team at the Plasma Physics Institut (IPP) in Garching has carried out a statistical analysis of the 5D distribution functions obtained from gyrokinetic tokamak simulations [38]. They discovered that the fluctuations are much higher in the space directions than in the velocity directions (see Figure 1).

This indicates that the approximation of the distribution function could require fewer data while still achieving a good representation, even in the collisionless regime.

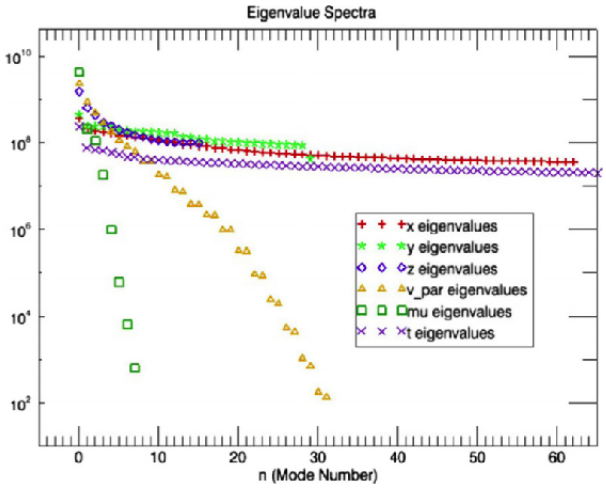


Figure 1. Space and velocity fluctuations spectra (from [38])

Our approach is different from the fluid approximation. In what follows we call this the "reduced model" approach. A reduced model is a model where the explicit dependence on the velocity variable is removed. In a more mathematical way, we consider that in some regions of the plasma, it is possible to exhibit a (preferably small) set of parameters α that allows us to describe the main properties of the plasma with a generalized "Maxwellian" M. Then

$$f(x, v, t) = M(\alpha(x, t), v).$$

In this case it is sufficient to solve for $\alpha(x, t)$. Generally, the vector α is solution of a first order hyperbolic system.

Several approaches are possible: waterbag approximations, velocity space transforms, *etc.*

3.2.1. Velocity space transformations

An experiment made in the 60's [41] exhibits in a spectacular way the reversible nature of the Vlasov equations. When two perturbations are applied to a plasma at different times, at first the plasma seems to damp and reach an equilibrium. But the information of the perturbations is still here and "hidden" in the high frequency microscopic oscillations of the distribution function. At a later time a resonance occurs and the plasma produces an echo. The time at which the echo occurs can be computed (see Villani [0], page 74). The fine mathematical study of this phenomenon allowed C. Villani and C. Mouhot to prove their famous result on the rigorous nonlinear Landau damping [42].

More practically, this experiment and its theoretical framework show that it is interesting to represent the distribution function by an expansion on an orthonormal basis of oscillating functions in the velocity variables. This representation allows a better control of the energy transfer between the low frequencies and the high frequencies in the velocity direction, and thus provides more relevant numerical methods. This kind of approach is studied for instance by Eliasson in [34] with the Fourier expansion.

[0] Landau damping. CEMRACS 2010 lectures. http://smai.emath.fr/cemracs/cemracs10/PROJ/Villani-lectures.pdf

In long time scales, filamentation phenomena result in high frequency oscillations in velocity space that numerical schemes cannot resolve. For stability purposes, most numerical schemes contain dissipation mechanisms that may affect the precision of the finest oscillations that could be resolved.

3.2.2. Adaptive modeling

Another trend in scientific computing is to optimize the computation time through adaptive modeling. This approach consists in applying the more efficient model locally, in the computational domain, according to an error indicator. In tokamak simulations, this kind of approach could be very efficient, if we are able to choose locally the best intermediate kinetic-fluid model as the computation runs. This field of research is very promising. It requires developing a clever hierarchy of models, rigorous error indicators, versatile software architecture, and algorithms adapted to new multicore computers.

3.2.3. Numerical schemes

As previously indicated, an efficient method for solving the reduced models is the Discontinuous Galerkin (DG) approach. It is possible to make it of arbitrary order. It requires limiters when it is applied to nonlinear PDEs occurring for instance in fluid mechanics. But the reduced models that we intent to write are essentially linear. The nonlinearity is concentrated in a few coupling source terms.

In addition, this method, when written in special set of variables, called the entropy variables, has nice properties concerning the entropy dissipation of the model. It opens the door to constructing numerical schemes with good conservation properties and no entropy dissipation, as already used for other systems of PDEs [44], [30], [40], [39].

3.3. Electromagnetic solvers

A precise resolution of the electromagnetic fields is essential for proper plasma simulation. Thus it is important to use efficient solvers for the Maxwell systems and its asymptotics: Poisson equation and magnetostatics.

The proper coupling of the electromagnetic solver with the Vlasov solver is also crucial for ensuring conservation properties and stability of the simulation.

Finally plasma physics implies very different time scales. It is thus very important to develop implicit Maxwell solvers and Asymptotic Preserving (AP) schemes in order to obtain good behavior on long time scales.

3.3.1. Coupling

The coupling of the Maxwell equations to the Vlasov solver requires some precautions. The most important is to control the charge conservation errors, which are related to the divergence conditions on the electric and magnetic fields. We will generally use divergence correction tools for hyperbolic systems presented for instance in [29] (and included references).

3.3.2. Implicit solvers

As already pointed out, in a tokamak, the plasma presents several different space and time scales. It is not possible in practice to solve the initial Vlasov-Maxwell model. It is first necessary to establish asymptotic models by letting some parameters (such as the Larmor frequency or the speed of light) tend to infinity. This is the case for the electromagnetic solver and this requires implementing implicit time solvers in order to efficiently capture the stationary state, the solution of the magnetic induction equation or the Poisson equation.

VIRTUAL PLANTS Project-Team

3. Research Program

3.1. Analysis of structures resulting from meristem activity

To analyze plant growth and structure, we focus mainly on methods for analyzing sequences and tree-structured data. Theses methods range from algorithms for computing distance between sequences or tree-structured data to statistical models.

- *Combinatorial approaches*: plant structures exhibit complex branching organizations of their organs like internodes, leaves, shoots, axes, branches, etc. These structures can be analyzed with combinatorial methods in order to compare them or to reveal particular types of organization. We investigate a family of techniques to quantify distances between branching systems based on non-linear structural alignment (similar to edit-operation methods used for sequence comparison). Based on these techniques, we study the notion of (topology-based) self-similarity of branching structures in order to define a notion of degree of redundancy for any tree structure and to quantify in this way botanical notions, such as the physiological states of a meristem, fundamental to the description of plant morphogenesis.
- *Statistical modeling*: We investigate different categories of statistical models corresponding to different types of structures.
  - Longitudinal data corresponding to plant growth follow up: the statistical models of interest are equilibrium renewal processes and generalized linear mixed models for longitudinal count data.
  - Repeated patterns within sequences or trees: the statistical models of interest are mainly (hidden) variable-order Markov chains. Hidden variable-order Markov chains were in particular applied to characterize permutation patterns in phyllotaxis and the alternation between flowering and vegetative growth units along sympodial tree axes.
  - Homogeneous zones (or change points) within sequences or trees: most of the statistical models of interest are hidden Markovian models (hidden semi-Markov chains, semi-Markov switching linear mixed models and semi-Markov switching generalized linear models for sequences and different families of hidden Markov tree models). A complementary approach consists in applying multiple change-point models. The branching structure of a parent shoot is often organized as a succession of branching zones while the succession of shoot at the more macroscopic scale exhibit roughly stationary phases separated by marked change points.

  We investigate both estimation methods and diagnostic tools for these different categories of models. In particular we focus on diagnostic tools for latent structure models (e.g. hidden Markovian models or multiple change-point models) that consist in exploring the latent structure space.
- *A new generation of morphogenesis models*: Designing morphogenesis models of the plant development at the macroscopic scales is a challenging problem. As opposed to modeling approaches that attempt to describe plant development on the basis of the integration of purely mechanistic models of various plant functions, we intend to design models that tightly couple mechanistic and empirical sub-models that are elaborated in our plant architecture analysis approach. Empirical models are used as a powerful complementary source of knowledge in places where knowledge about mechanistic processes is lacking or weak. We chose to implement such integrated models in a programming language dedicated to dynamical systems with dynamical structure $(DS)^2$, such as L-systems or MGS. This type of language plays the role of an integration framework for sub-models of heterogeneous nature.

3.2. Meristem functioning and development

In this second scientific axis, we develop models of meristem growth at tissue level in order to integrate various sources of knowledge and to analyze their dynamic and complex spatial interaction. To carry out this integration, we need to develop a complete methodological approach containing:

- algorithms for the automatized segmentation in 3D, and cell lineage tracking throughout time, for images coming from confocal microscopy,
- design of high-level routines and user interfaces to distribute these image analysis tools to the scientific community,
- tools for structural and statistical analysis of 3D meristem structure (spatial statistics, multiscale geometric and topological analysis),
- physical models of cells interactions based on spring-mass systems or on tensorial mechanics at the level of cells,
- models of biochemical networks of hormonal and gene driven regulation, at the cellular and tissue level, using continuous and discrete formalisms,
- and models of cell development taking into account the effects of growth and cell divisions on the two previous classes of models.

3.3. OpenAlea: An open-software platform for plant modeling

OpenAlea is open-software platform for interdisciplinary research in plant modeling and simulation. This scientific workflow platform is used for the integration and comparison of different models and tools provided by the research community. It is based on the Python (http://www.python.org) language that aims at being both a *glue* language for the different modules and an efficient modeling language for developing new models and tools. *OpenAlea* currently includes modules for plant simulation, analysis and modeling at different scales (*V-Plants* modules), for modeling ecophysiological processes (*Alinea* modules) such as radiative transfer, transpiration and photosynthesis (*RATP*, *Caribu*, *Adel*, *TopVine*, *Ecomeristem*) and for 3D visualization of plant architecture at different scales (*PlantGL*).

OpenAlea is the result of a collaborative effort associating 20 french research teams in plant modeling from Inria, CIRAD, INRA and ENS Lyon. The Virtual Plants team coordinates both development and modeling consortiums, and is more particularly in charge of the development of the kernel and of some of the main data structures such as multi-scale tree graphs and statistical sequences.

OpenAlea is a fundamental tool to share models and methods in interdisciplinary research (comprising botany, ecophysiology, forestry, agronomy, applied mathematics and computer science approaches). Embedded in Python and its scientific libraries, the platform may be used as a flexible and useful toolbox by biologists and modelers for various purposes (research, teaching, rapid model prototyping, communication, etc.).

VISAGES Project-Team

3. Research Program

3.1. Research Program

The scientific foundations of our team concern the development of new processing algorithms in the field of medical image computing : image fusion (registration and visualization), image segmentation and analysis, management of image related information. Since this is a very large domain, which can endorse numerous types of application; for seek of efficiency, the purpose of our methodological work primarily focuses on clinical aspects and for the most part on head and neck related diseases. In addition, we emphasize our research efforts on the neuroimaging domain. Concerning the scientific foundations, we have pushed our research efforts:

- In the field of image fusion and image registration (rigid and deformable transformations) with a special emphasis on new challenging registration issues, especially when statistical approaches based on joint histogram cannot be used or when the registration stage has to cope with loss or appearance of material (like in surgery or in tumor imaging for instance).
- In the field of image analysis and statistical modeling with a new focus on image feature and group analysis problems. A special attention was also to develop advanced frameworks for the construction of atlases and for automatic and supervised labeling of brain structures.
- In the field of image segmentation and structure recognition, with a special emphasis on the difficult problems of *i*) image restoration for new imaging sequences (new Magnetic Resonance Imaging protocols, 3D ultrasound sequences...), and *ii*) structure segmentation and labelling based on shape, multimodal and statistical information.
- Following the Neurobase national project where we had a leading role, we wanted to enhance the development of distributed and heterogeneous medical image processing systems.

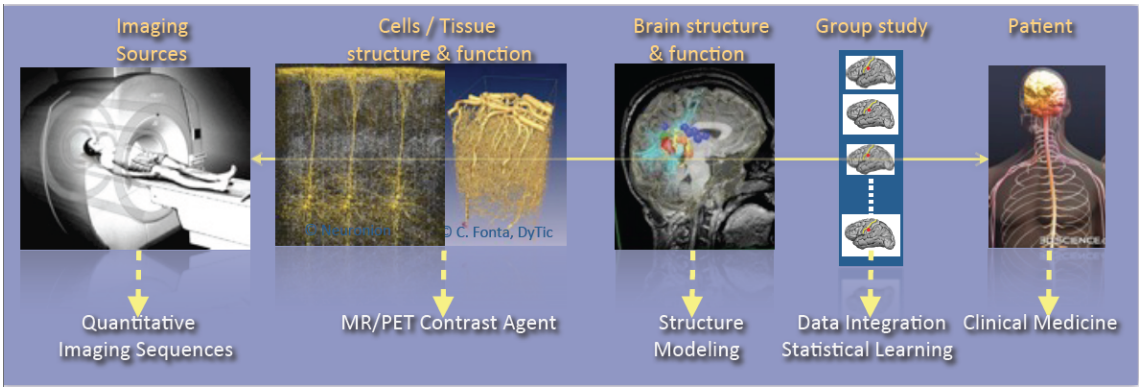


Figure 1. The major overall scientific foundation of the team concerns the integration of data from the Imaging source to the patient at different scales : from the cellular or molecular level describing the structure and function, to the functional and structural level of brain structures and regions, to the population level for the modelling of group patterns and the learning of group or individual imaging markers

As shown in figure 1 , research activities of the VISAGES U746 team are tightly coupling observations and models through integration of clinical and multi-scale data, phenotypes (cellular, molecular or structural patterns). We work on personalized models of central nervous system organs and pathologies, and intend to confront these models to clinical investigation studies for quantitative diagnosis, prevention of diseases, therapy planning and validation. These approaches are developed in a translational framework where the data integration process to build the models inherits from specific clinical studies, and where the models are assessed on prospective clinical trials for diagnosis and therapy planning. All of this research activity is conducted in tight links with the Neurinfo imaging platform environments and the engineering staff of the platform. In this context, some of our major challenges in this domain concern:

- The elaboration of new descriptors to study the brain structure and function (e.g. variation of brain perfusion with and without contrast agent, evolution in shape and size of an anatomical structure in relation with normal, pathological or functional patterns, computation of asymmetries from shapes and volumes).
- The integration of additional spatio-temporal imaging sequences covering a larger range of observation, from the molecular level to the organ through the cell (Arterial Spin Labeling, diffusion MRI, MR relaxometry, MR cell labeling imaging, PET molecular imaging, ...). This includes the elaboration of new image descriptors coming from spatio-temporal quantitative or contrast-enhanced MRI.
- The creation of computational models through data fusion of molecular, cellular, structural and functional image descriptors from group studies of normal and/or pathological subjects.
- The evaluation of these models on acute pathologies especially for the study of degenerative, psychiatric or developmental brain diseases (e.g. Multiple Sclerosis, Epilepsy, Parkinson, Dementia, Strokes, Depression, Schizophrenia, ...) in a translational framework.

In terms of methodological developments, we are particularly working on statistical methods for multidimensional image analysis, and feature selection and discovery, which includes:

- The development of specific shape and appearance models, construction of atlases better adapted to a patient or a group of patients in order to better characterize the pathology;
- The development of advanced segmentation and modeling methods dealing with longitudinal and multidimensional data (vector or tensor fields), especially with the integration of new prior models to control the integration of multiscale data and aggregation of models;
- The development of new models and probabilistic methods to create water diffusion maps from MRI;
- The integration of machine learning procedures for classification and labeling of multidimensional features (from scalar to tensor fields and/or geometric features): pattern and rule inference and knowledge extraction are key techniques to help in the elaboration of knowledge in the complex domains we address;
- The development of new dimensionality reduction techniques for problems with massive data, which includes dictionary learning for sparse model discovery. Efficient techniques have still to be developed to properly extract from a raw mass of images derived data that are easier to analyze.

ALGORILLE Project-Team

3. Research Program

3.1. Structuring Applications

Computing on different scales is a challenge under constant development that, almost by definition, will always try to reach the edge of what is possible at any given moment in time: in terms of the scale of the applications under consideration, in terms of the efficiency of implementations and in what concerns the optimized utilization of the resources that modern platforms provide or require. The complexity of all these aspects is currently increasing rapidly.

3.1.1. Diversity of platforms

Design of processing hardware is diverging in many different directions. Nowadays we have SIMD registers inside processors, on-chip or off-chip accelerators (many-core boards, GPU, FPGA, vector-units), multi-cores and hyperthreading, multi-socket architectures, clusters, grids, clouds... The classical monolithic architecture of one-algorithm/one-implementation that solves a problem is obsolete in many cases. Algorithms (and the software that implements them) must deal with this variety of execution platforms robustly.

As we know, the "*free lunch*" for sequential algorithms provided by the increase of processor frequencies is over, we have to go parallel. But the "*free lunch*" is also over for many automatic or implicit adaptation strategies between codes and platforms: e.g the best cache strategies can't help applications that access memory randomly, or algorithms written for "simple" CPU (von Neumann model) have to be adapted substantially to run efficiently on vector units.

3.1.2. The communication bottleneck

Communication and processing capacities evolve at a different pace, thus the *communication bottleneck* is always narrowing. An efficient data management is becoming more and more crucial.

Not many implicit data models have yet found their place in the HPC domain, because of a simple observation: latency issues easily kill the performance of such tools. In the best case, they will be able to hide latency by doing some intelligent caching and delayed updating. But they can never hide the bottleneck for bandwidth. An efficient solution to this problem is the use of asynchronism in the algorithms. However, until now its application has been limited to iterative processes with specific constraints over the computational scheme.

HPC was previously able to cope with the communication bottleneck by using an explicit model of communication, namely MPI. It has the advantage of imposing explicit points in code where some guarantees about the state of data can be given. It has the clear disadvantage that coherence of data between different participants is difficult to manage and is completely left to the programmer.

Here, our approach is and will be to timely request explicit actions (like MPI) that mark the availability of (or need for) data. Such explicit actions ease the coordination between tasks (coherence management) and allow the platform underneath the program to perform a pro-active resource management.

3.1.3. Models of interdependence and consistency

Interdependence of data between different tasks of an application and components of hardware will be crucial to ensure that developments will possibly scale on the ever diverging architectures. We have up to now presented such models (PRO, DHO, ORWL) and their implementations, and proved their validity for the context of SPMD-type algorithms.

Over the next years we will have to enlarge the spectrum of their application. On the algorithm side we will have to move to heterogeneous computations combining different types of tasks in one application. Concerning the architectures, we will have to take into account the fact of increased heterogeneity, processors of different speeds, multi-cores, accelerators (FPU, GPU, vector units), communication links of different bandwidth and latency, memory and generally storage capacity of different size, speed and access characteristics. First implementations using ORWL in that context look particularly promising.

The models themselves will have to evolve to be better suited for more types of applications, such that they allow for a more fine-grained partial locking and access of objects. They should handle *e.g.* collaborative editing or the modification of just some fields in a data structure. This work has already started with DHO which allows the locking of *data ranges* inside an object. But a more structured approach would certainly be necessary here to be usable more comfortably in most applications.

3.1.4. Frequent I/O

A complete parallel application includes I/O of massive data, at an increasing frequency. In addition to applicative input and output data flows, I/O are used for checkpointing or to store traces of execution. These then can be used to restart in case of failure (hardware or software) or for a post-mortem analysis of a chain of computations that led to catastrophic actions (for example in finance or in industrial system control). The difficulty of frequent I/O is more pronounced on hierarchical parallel architectures that include accelerators with local memory.

I/O have to be included in the design of parallel programming models and tools. The ORWL library (Ordered Read-Write Lock) should be enriched with such tools and functionalities, in order to ease the modeling and development of parallel applications that include data IO, and to exploit most of the performance potential of parallel and distributed architectures.

3.1.5. Algorithmic paradigms

Concerning asynchronous algorithms, we have studied different variants of asynchronous models and developed several versions of implementations, allowing us to precisely study the impact of our design choices. However, we are still convinced that improvements are possible in order to extend the applicability of asynchronism, especially concerning the control of its behavior and the termination detection (global convergence of iterative algorithms). We have proposed some generic and non-intrusive way of implementing such a procedure in any parallel iterative algorithm.

3.1.6. Cost models and accelerators

We have already designed some models that relate computation power and energy consumption. Our present works in this topic concern the design and implementation of an auto-tuning system that controls the application according to user defined optimization criteria (computation and/or energy performance). This implies the insertion of multi-schemes and/or multi-kernels into the application such that it will be able to adapt its behavior to the requirements.

3.1.7. Design of dynamical systems for computational tasks

In the context of a collaboration with Nazim Fatès over dynamical systems, and especially cellular automata, we address a new way to study dynamical systems, that is more development oriented than analysis oriented. In fact, until now, most of the studies related to dynamical systems consisted in analyzing the dynamical properties (convergence, fixed points, cycles, initialization,...) of some given systems, and in describing the emergence of complex behaviors. Here, we focus on the dual approach that consists in designing dynamical systems in order to fulfill some given tasks. In this approach, we consider both theoretical and practical aspects.

3.2. Transparent Resource Management for Clouds

Given the extremely large offer of resources by public or private clouds, users need software assistance to make provisioning decisions. Our goal is to design a **cloud resource broker** which handles the workload of a user or

of a community of users as a multi-criteria optimization problem. The notions of resource usage, scheduling, provisioning and task management have been adapted to this new context. For example, to minimize the makespan of a DAG of tasks, usually a fixed number of resources is assumed. On IaaS clouds, the amount of resources can be provisioned at any time, and hence the scheduling problem must be redefined using one new prevalent optimization criterion: the financial cost of the computation.

3.2.1. Provisioning strategies

The provisioning strategies are hence central to the broker. They are designed after heuristics which aim to fit execution constraints and satisfy user preferences. For instance, lowering the costs can be achieved with strategies aiming at reusing already leased resources, or switch to less powerful and cheaper resources. However, some economic models proposed by cloud providers involve a complex cost-benefit analysis which we plan to address. Moreover, these economic models incur additional costs, *e.g.* for data storage or transfer, which have to be taken into account to design a comprehensive broker.

3.2.2. User workload analysis

Another possible extension of the capability of such a broker is the analysis of user workloads. Characterizing the workload might help to anticipate the behavior of each alternative provisioning strategy. The objective is to allow the user to select the suitable provisioning solution thanks to concrete information, such as completion time and financial cost.

3.2.3. Simulation of cloud platforms

Providing concrete information about provisioning solutions can also be achieved through simulation. Although predicting the behavior of applicative cases in real grid environment is made very difficult by the shared (*e.g.* multi-tenant), heterogeneous and dynamic nature of the resources, cloud resources (i.e. VMs) are perceived as reserved and homogeneous and stable by the end-user. Therefore, proposing an accurate prediction of the different strategies through an accurate simulation process would be a strong decision support for the user.

3.3. Experimental Methodologies for the Evaluation of Distributed Systems

Distributed systems are very challenging to study, test, and evaluate. Computer scientists traditionally prefer to study their systems *a priori* by reasoning theoretically on the constituents and their interactions. But the complexity of large-scale distributed systems makes this methodology near to impossible, explaining that most of the studies are done *a posteriori* through experiments.

In AlGorille, we strive at designing a comprehensive set of solutions for experimentation on distributed systems by working on several methodologies (formal assessment, simulation, use of experimental facilities, emulation) and by leveraging the convergence opportunities between methodologies (co-development, shared interfaces, validation combining several methodologies).

3.3.1. Simulation and Dynamic Verification

Our team plays a key role in the SimGrid project, a mature simulation toolkit widely used in the distributed computing community. Since more than ten years, we work on the validity, scalability and robustness of our tool.

Our current medium term goal is to extend the tool applicability to **Clouds and Exascale systems**. In the last years, we therefore worked toward disk and memory models in addition to the previously existing network and CPU models. The tool's scalability and efficiency also constitutes a permanent concern to us. **Interfaces** constitute another important work axis, with the addition of specific APIs on top of our simulation kernel. They provide the "syntactic sugar" needed to express algorithms of these communities. For example, virtual machines are handled explicitly in the interface provided for Cloud studies. Similarly, we pursue our work on an implementation of the full MPI standard allowing to study real applications using that interface. This work may also be extended in the future to other interfaces such as OpenMP or OpenCL.

We integrated a model checking kernel in SimGrid to enable **formal correctness studies** in addition to the practical performance studies enabled by simulation. Being able to study these two fundamental aspects of distributed applications within the same tool constitutes a major advantage for our users. In the future, we will enforce this capacity for the study of correctness and performance such that we hope to tackle their usage on real applications.

3.3.2. Experimentation on testbeds and production facilities, emulation

Our work in this research axis is meant to bring major contributions to the **industrialization of experimentation** on parallel and distributed systems. It is structured through multiple layers that range from the design of a testbed supporting high-quality experimentation, to the study of how stringent experimental methodology could be applied to our field, as depicted in Figure 2 .

During the last years, we have played a **key role in the design and development of Grid'5000** by leading the design and technical developments, and by managing several engineers working on the platform. We pursue our involvement in the design of the testbed with a focus on ensuring that the testbed provides all the features needed for high-quality experimentation. We also collaborate with other testbeds sharing similar goals in order to exchange ideas and views. We now work on **basic services supporting experimentation** such as resources verification, management of experimental environments, control of nodes, management of data, etc. Appropriate collaborations will ensure that existing solutions are adopted to the platform and improved as much as possible.

One key service for experimentation is the ability to alter experimental conditions using emulation. We work on the **Distem emulator**, focusing on its validation and on adding features (such as the ability to emulate faults, varying availability, churn, load injection, etc) and investigate if altering memory and disk performance is possible. Other goals are to scale the tool up to 20000 virtual nodes while improving the tool usability and documentation.

We work on **orchestration of experiments** in order to combine all the basic services mentioned previously in an efficient and scalable manner, with the design of a workflow-based experiment control engine named **XPFlow**.

3.3.3. Convergence and co-design of experimental methodologies

We see the experimental methodologies we work on as steps of a common experimental staircase: ideally, **one could and should leverage the various methodologies to address different facets of the same problem**. To facilitate that, we must co-design common or compatible formalisms, semantics and data formats.

Other experimental sciences such as biology and physics have paved the way in terms of scientific methodology. We **should learn from other experimental sciences, adopt good practices and adapt them** to Computer Science's specificities.

But Computer Science also has specific features that make it the ideal field to **create a truly Open Science**: provide infrastructure and tools for publishing and reproducing experiments and results, linked with our own methodologies and tools.

Finally, one important part of our work is to maintain a deep understanding of systems and their environments, in order to properly model them and experiment on them. Similarly, we need to understand the emerging scientific challenges in our field in order to improve adequately our experimental tools.

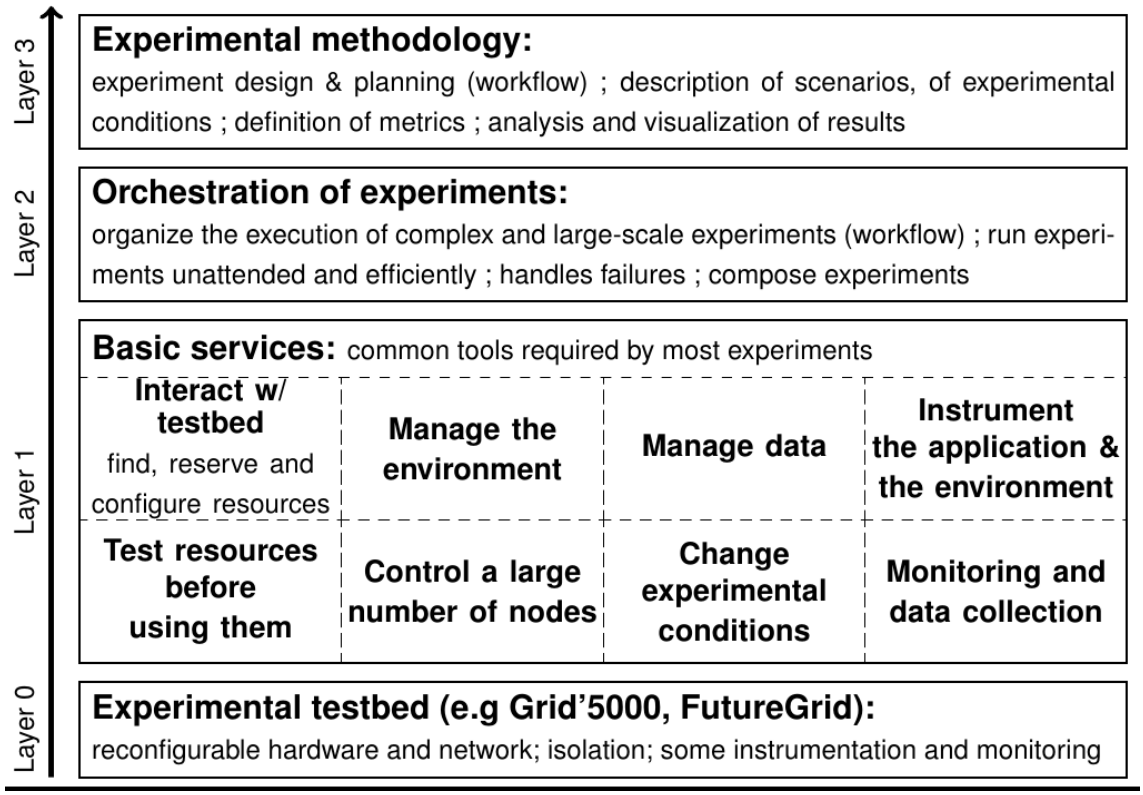


Figure 2. General structure of our project: We plan to address all layers of the experimentation stack.

ALPINES Project-Team

3. Research Program

3.1. Overview

The research described here is directly relevant to several steps of the numerical simulation chain. Given a numerical simulation that was expressed as a set of differential equations, our research focuses on mesh generation methods for parallel computation, novel numerical algorithms for linear algebra, as well as algorithms and tools for their efficient and scalable implementation on high performance computers. The validation and the exploitation of the results will be performed with collaborators from applications and it will be based on the usage of existing tools. In summary, the topics studied in our group are the following:

- Numerical methods and algorithms
  - Mesh generation for parallel computation
  - Solvers for numerical linear algebra
  - Computational kernels for numerical linear algebra
- Validation on numerical simulations

3.2. Domain specific language - parallel FreeFem++

In the engineering, researchers, and teachers communities, there is a strong demand for simulation frameworks that are simple to install and use, efficient, sustainable, and that solve efficiently and accurately complex problems for which there are no dedicated tools or codes available. In our group we develop FreeFem++ (see http://www.freefem.org/ff++), a user dedicated language for solving PDEs. The goal of FreeFem++ is not to be a substitute for complex numerical codes, but rather to provide an efficient and relatively generic tool for:

- getting a quick answer to a specific problem,
- prototype the resolution of a new complex problem.

The current users of FreeFem++ are mathematicians, engineers, university professors, and students. In general for these users the installation of public libraries as MPI, MUMPS, Ipopt, Blas, lapack, OpenGL, fftw, scotch, is a very difficult problem. For this reason, the authors of FreeFem++ have created a user friendly language, and over years have enriched its capabilities and provided tools for compiling FreeFem++ such that the users do not need to have special knowledge of computer science. This leads to an important work on porting the software on different emerging architectures.

Today, the main components of parallel FreeFem++ are:

1. definition of a coarse grid,
2. splitting of the coarse grid,
3. mesh generation of all subdomains of the coarse grid, and construction of parallel datat structures for vectors and sparse matrices from the mesh of the subdomain,
4. call to a linear solver,
5. analysis of the result.

All these components are parallel, except for point (5) which is not in the focus of our research. However for the moment, the parallel mesh generation algorithm is very simple and not sufficient, for example it addresses only polygonal geometries. Having a better parallel mesh generation algorithm is one of the goals of our project. In addition, in the current version of FreeFem++, the parallelism is not hidden from the user, it is done through direct calls to MPI. Our goal is also to hide all the MPI calls in the specific language part of FreeFem++.

3.3. Solvers for numerical linear algebra

Iterative methods are widely used in industrial applications, and preconditioning is the most important research subject here. Our research considers domain decomposition methods and iterative methods and its goal is to develop solvers that are suitable for parallelism and that exploit the fact that the matrices are arising from the discretization of a system of PDEs on unstructured grids.

One of the main challenges that we address is the lack of robustness and scalability of existing methods as incomplete LU factorizations or Schwarz-based approaches, for which the number of iterations increases significantly with the problem size or with the number of processors. This is often due to the presence of several low frequency modes that hinder the convergence of the iterative method. To address this problem, we study direction preserving solvers in the context of multilevel domain decomposition methods with adaptive coarse spaces and multilevel incomplete decompositions. A judicious choice for the directions to be preserved through filtering or low rank approximations allows us to alleviate the effect of low frequency modes on the convergence.

We also focus on developing boundary integral equation methods that would be adapted to the simulation of wave propagation in complex physical situations, and that would lend themselves to the use of parallel architectures, which includes devising adapted domain decomposition approaches. The final objective is to bring the state of the art on boundary integral equations closer to contemporary industrial needs.

3.4. Computational kernels for numerical linear algebra

The design of new numerical methods that are robust and that have well proven convergence properties is one of the challenges addressed in Alpines. Another important challenge is the design of parallel algorithms for the novel numerical methods and the underlying building blocks from numerical linear algebra. The goal is to enable their efficient execution on a diverse set of node architectures and their scaling to emerging high-performance clusters with an increasing number of nodes.

Increased communication cost is one of the main challenges in high performance computing that we address in our research by investigating algorithms that minimize communication, as communication avoiding algorithms. We propose to integrate the minimization of communication into the algorithmic design of numerical linear algebra problems. This is different from previous approaches where the communication problem was addressed as a scheduling or as a tuning problem. The communication avoiding algorithmic design is an aproach originally developed in our group since 2007 (initially in collaboration with researchers from UC Berkeley and CU Denver). While at mid term we focus on reducing communication in numerical linear algebra, at long term we aim at considering the communication problem one level higher, during the parallel mesh generation tool described earlier.

ASAP Project-Team

3. Research Program

3.1. Distributed computing

Distributed computing was born in the late seventies when people started taking into account the intrinsic characteristics of physically distributed systems. The field then emerged as a specialized research area distinct from networks, operating systems and parallelism. Its birth certificate is usually considered as the publication in 1978 of Lamport's most celebrated paper "*Time, clocks and the ordering of events in a distributed system*" [55] (that paper was awarded the Dijkstra Prize in 2000). Since then, several high-level journals and (mainly ACM and IEEE) conferences have been devoted to distributed computing. The distributed systems area has continuously been evolving, following the progresses of all the above-mentioned areas such as networks, computing architecture, operating systems.

The last decade has witnessed significant changes in the area of distributed computing. This has been acknowledged by the creation of several conferences such as NSDI and IEEE P2P. The NSDI conference is an attempt to reassemble the networking and system communities while the IEEE P2P conference was created to be a forum specialized in peer-to-peer systems. At the same time, the EuroSys conference originated as an initiative of the European Chapter of the ACM SIGOPS to gather the system community in Europe.

3.2. Theory of distributed systems

Finding models for distributed computations prone to asynchrony and failures has received a lot of attention. A lot of research in this domain focuses on what can be computed in such models, and, when a problem can be solved, what are its best solutions in terms of relevant cost criteria. An important part of that research is focused on distributed computability: what can be computed when failure detectors are combined with conditions on process input values for example. Another part is devoted to model equivalence. What can be computed with a given class of failure detectors? Which synchronization primitives is a given failure class equivalent to? These are among the main topics addressed in the leading distributed computing community. A second fundamental issue related to distributed models, is the definition of appropriate models suited to dynamic systems. Up to now, the researchers in that area consider that nodes can enter and leave the system, but do not provide a simple characterization, based on properties of computation instead of description of possible behaviors [56], [50], [51]. This shows that finding dynamic distributed computing models is today a "Holy Grail", whose discovery would allow a better understanding of the essential nature of dynamic systems.

3.3. Peer-to-peer overlay networks

A standard distributed system today is related to thousands or even millions of computing entities scattered all over the world and dealing with a huge amount of data. This major shift in scalability requirements has lead to the emergence of novel computing paradigms. In particular, the peer-to-peer communication paradigm imposed itself as the prevalent model to cope with the requirements of large scale distributed systems. Peer-to-peer systems rely on a symmetric communication model where peers are potentially both clients and servers. They are fully decentralized, thus avoiding the bottleneck imposed by the presence of servers in traditional systems. They are highly resilient to peers arrivals and departures. Finally, individual peer behavior is based on a local knowledge of the system and yet the system converges toward global properties.

A peer-to-peer overlay network logically connects peers on top of IP. Two main classes of such overlays dominate, structured and unstructured. The differences relate to the choice of the neighbors in the overlay, and the presence of an underlying naming structure. Overlay networks represent the main approach to build large-scale distributed systems that we retained. An overlay network forms a logical structure connecting participating entities on top of the physical network, be it IP or a wireless network. Such an overlay might form a structured overlay network [57], [58], [59] following a specific topology or an unstructured network [54], [60] where participating entities are connected in a random or pseudo-random fashion. In between, lie weakly structured peer-to-peer overlays where nodes are linked depending on a proximity measure providing more flexibility than structured overlays and better performance than fully unstructured ones. Proximity-aware overlays connect participating entities so that they are connected to close neighbors according to a given proximity metric reflecting some degree of affinity (computation, interest, etc.) between peers. We extensively use this approach to provide algorithmic foundations of large-scale dynamic systems.

3.4. Epidemic protocols

Epidemic algorithms, also called gossip-based algorithms [53], [52], constitute a fundamental topic in our research. In the context of distributed systems, epidemic protocols are mainly used to create overlay networks and to ensure a reliable information dissemination in a large-scale distributed system. The principle underlying technique, in analogy with the spread of a rumor among humans via gossiping, is that participating entities continuously exchange information about the system in order to spread it gradually and reliably. Epidemic algorithms have proved efficient to build and maintain large-scale distributed systems in the context of many applications such as broadcasting [52], monitoring, resource management, search, and more generally in building unstructured peer-to-peer networks.

3.5. Malicious process behaviors

When assuming that processes fail by simply crashing, bounds on resiliency (maximum number of processes that may crash), number of exchanged messages, number of communication steps, etc. either in synchronous and augmented asynchronous systems (recall that in purely asynchronous systems some problems are impossible to solve) are known. If processes can exhibit malicious behaviors, these bounds are seldom the same. Sometimes, it is even necessary to change the specification of the problem. For example, the consensus problem for correct processes does not make sense if some processes can exhibit a Byzantine behavior and thus propose an arbitrary value. In this case, the validity property of consensus, which is normally "a decided value is a proposed value", must be changed to "if all correct processes propose the same value then only this value can be decided." Moreover, the resilience bound of less than half of faulty processes is at least lowered to "less than a third of Byzantine processes." These are some of the aspects that underlie our studies in the context of the classical model of distributed systems, in peer-to-peer systems and in sensor networks.

3.6. Online social networks and recommender systems

Social Networks have rapidly become a fundamental component of today's distributed applications. Web 2.0 applications have dramatically changed the way users interact with the Internet and with each other. The number of users of websites like Flickr, Delicious, Facebook, or MySpace is constantly growing, leading to significant technical challenges. On the one hand, these websites are called to handle enormous amounts of data. On the other hand, news continue to report the emergence of privacy threats to the personal data of social-network users. Our research aims to exploit our expertise in distributed systems to lead to a new generation of scalable, privacy-preserving, social applications.

We also investigate approaches to build implicit social networks, connecting users sharing similar interests. At the heart of the building of such similarity graphs lie k-nearest neighbor (KNN) algorithms. Our research in this area is to design and implement efficient KNN algorithms able to cope with a huge volume of data as well as a high level of dynamism. We investigate the use of such similarity graphs to build highly scalable infrastructures for recommendation systems.

ASCOLA Project-Team

3. Research Program

3.1. Overview

Since we mainly work on new concepts for the language-based definition and implementation of complex software systems, we first briefly introduce some basic notions and problems of software components (understood in a broad sense, that is, including modules, objects, architecture description languages and services), aspects, and domain-specific languages. We conclude by presenting the main issues related to distribution and concurrency, in particular related to capacity planning issues that are relevant to our work.

3.2. Software Composition

Modules and services. The idea that building *software components*, i.e., composable prefabricated and parameterized software parts, was key to create an effective software industry was realized very early [64]. At that time, the scope of a component was limited to a single procedure. In the seventies, the growing complexity of software made it necessary to consider a new level of structuring and programming and led to the notions of information hiding, *modules*, and module interconnection languages [71], [49]. Information hiding promotes a black-box model of program development whereby a module implementation, basically a collection of procedures, is strongly encapsulated behind an interface. This makes it possible to guarantee logical invariant *properties* of the data managed by the procedures and, more generally, makes *modular reasoning* possible.

In the context of today's Internet-based information society, components and modules have given rise to *software services* whose compositions are governed by explicit *orchestration or choreography* specifications that support notions of global properties of a service-oriented architecture. These horizontal compositions have, however, to be frequently adapted dynamically. Dynamic adaptations, in particular in the context of software evolution processes, often conflict with a black-box composition model either because of the need for invasive modifications, for instance, in order to optimize resource utilization or modifications to the vertical compositions implementing the high-level services.

Object-Oriented Programming. Classes and objects provide another kind of software component, which makes it necessary to distinguish between *component types* (classes) and *component instances* (objects). Indeed, unlike modules, objects can be created dynamically. Although it is also possible to talk about classes in terms of interfaces and implementations, the encapsulation provided by classes is not as strong as the one provided by modules. This is because, through the use of inheritance, object-oriented languages put the emphasis on *incremental programming* to the detriment of modular programming. This introduces a white-box model of software development and more flexibility is traded for safety as demonstrated by the *fragile base class* issue [67].

Architecture Description Languages. The advent of distributed applications made it necessary to consider more sophisticated connections between the various building blocks of a system. The *software architecture* [75] of a software system describes the system as a composition of *components* and *connectors*, where the connectors capture the *interaction protocols* between the components [40]. It also describes the rationale behind such a given architecture, linking the properties required from the system to its implementation. *Architecture Description Languages* (ADLs) are languages that support architecture-based development [65]. A number of these languages make it possible to generate executable systems from architectural descriptions, provided implementations for the primitive components are available. However, guaranteeing that the implementation conforms to the architecture is an issue.

Protocols. Today, protocols constitute a frequently used means to precisely define, implement, and analyze contracts, notably concerning communication and security properties, between two or more hardware or software entities. They have been used to define interactions between communication layers, security properties of distributed communications, interactions between objects and components, and business processes.

Object interactions [69], component interactions [81], [73] and service orchestrations [50] are most frequently expressed in terms of *regular interaction protocols* that enable basic properties, such as compatibility, substitutability, and deadlocks between components to be defined in terms of basic operations and closure properties of finite-state automata. Furthermore, such properties may be analyzed automatically using, e.g., model checking techniques [47], [56].

However, the limited expressive power of regular languages has led to a number of approaches using more expressive *non-regular* interaction protocols that often provide distribution-specific abstractions, e.g., session types [58], or context-free or turing-complete expressiveness [74], [45]. While these protocol types allow conformance between components to be defined (e.g., using unbounded counters), property verification can only be performed manually or semi-automatically.

3.3. Programming languages for advanced modularization

The main driving force for the structuring means, such as components and modules, is the quest for clean *separation of concerns* [51] on the architectural and programming levels. It has, however, early been noted that concern separation in the presence of crosscutting functionalities requires specific language and implementation level support. Techniques of so-called *computational reflection*, for instance, Smith's 3-Lisp or Kiczales's CLOS meta-object protocol [76], [61] as well as metaprogramming techniques have been developed to cope with this problem but proven unwieldy to use and not amenable to formalization and property analysis due to their generality. Methods and techniques from two fields have been particularly useful in addressing such advanced modularization problems: Aspect-Oriented Software Development as the field concerned with the systematic handling of modularization issues and domain-specific languages that provide declarative and efficient means for the definition of crosscutting functionalities.

Aspect-Oriented Software Development [60], [38] has emerged over the previous decade as the domain of systematic exploration of crosscutting concerns and corresponding support throughout the software development process. The corresponding research efforts have resulted, in particular, in the recognition of *crosscutting* as a fundamental problem of virtually any large-scale application, and the definition and implementation of a large number of aspect-oriented models and languages.

However, most current aspect-oriented models, notably AspectJ [59], rely on pointcuts and advice defined in terms of individual execution events. These models are subject to serious limitations concerning the modularization of crosscutting functionalities in distributed applications, the integration of aspects with other modularization mechanisms such as components, and the provision of correctness guarantees of the resulting AO applications. They do, in particular, only permit the manipulation of distributed applications on a per-host basis, that is, without direct expression of coordination properties relating different distributed entities [77]. Similarly, current approaches for the integration of aspects and (distributed) components do not directly express interaction properties between sets of components but rather seemingly unrelated modifications to individual components [48]. Finally, current formalizations of such aspect models are formulated in terms of low-level semantic abstractions (see, e.g., Wand's et al semantics for AspectJ [80]) and provide only limited support for the analysis of fundamental aspect properties.

Different approaches have been put forward to tackle these problems, in particular, in the context of so-called *stateful* or *history-based aspect languages* [52], [53], which provide pointcut and advice languages that directly express rich relationships between execution events. Such languages have been proposed to directly express coordination and synchronization issues of distributed and concurrent applications [70], [43], [55], provide more concise formal semantics for aspects and enable analysis of their properties [41], [54], [52], [39]. Furthermore, first approaches for the definition of *aspects over protocols* have been proposed, as well as over regular structures [52] and non-regular ones [79], [68], which are helpful for the modular definition and verification of protocols over crosscutting functionalities.

They represent, however, only first results and many important questions concerning these fundamental issues remain open, in particular, concerning the semantics foundations of AOP and the analysis and enforcement of correctness properties governing its, potentially highly invasive, modifications.

Domain-specific languages (DSLs) represent domain knowledge in terms of suitable basic language constructs and their compositions at the language level. By trading generality for abstraction, they enable complex relationships among domain concepts to be expressed concisely and their properties to be expressed and formally analyzed. DSLs have been applied to a large number of domains; they have been particularly popular in the domain of software generation and maintenance [66], [82].

Many modularization techniques and tasks can be naturally expressed by DSLs that are either specialized with respect to the type of modularization constructs, such as a specific brand of software component, or to the compositions that are admissible in the context of an application domain that is targeted by a modular implementation. Moreover, software development and evolution processes can frequently be expressed by transformations between applications implemented using different DSLs that represent an implementation at different abstraction levels or different parts of one application.

Functionalities that crosscut a component-based application, however, complicate such a DSL-based transformational software development process. Since such functionalities belong to another domain than that captured by the components, different DSLs should be composed. Such compositions (including their syntactic expression, semantics and property analysis) have only very partially been explored until now. Furthermore, restricted composition languages and many aspect languages that only match execution events of a specific domain (e.g., specific file accesses in the case of security functionality) and trigger only domain-specific actions clearly are quite similar to DSLs but remain to be explored.

3.4. Distribution and Concurrency

While ASCOLA does not investigate distribution and concurrency as research domains per se (but rather from a software engineering and modularization viewpoint), there are several specific problems and corresponding approaches in these domains that are directly related to its core interests that include the structuring and modularization of large-scale distributed infrastructures and applications. These problems include crosscutting functionalities of distributed and concurrent systems, support for the evolution of distributed software systems, and correctness guarantees for the resulting software systems.

Underlying our interest in these domains is the well-known observation that large-scale distributed applications are subject to *numerous crosscutting functionalities* (such as the transactional behavior in enterprise information systems, the implementation of security policies, and fault recovery strategies). These functionalities are typically partially encapsulated in distributed infrastructures and partially handled in an ad hoc manner by using infrastructure services at the application level. Support for a more principled approach to the development and evolution of distributed software systems in the presence of crosscutting functionalities has been investigated in the field of *open adaptable middleware* [44], [63]. Open middleware design exploits the concept of reflection to provide the desired level of configurability and openness. However, these approaches are subject to several fundamental problems. One important problem is their insufficient, framework-based support that only allows partial modularization of crosscutting functionalities.

There has been some *criticism* on the use of *AspectJ-like aspect models* (which middleware aspect models like that of JBoss AOP are an instance of) for the modularization of distribution and concurrency related concerns, in particular, for transaction concerns [62] and the modularization of the distribution concern itself [77]. Both criticisms are essentially grounded in AspectJ's inability to explicitly represent sophisticated relationships between execution events in a distributed system: such aspects therefore cannot capture the semantic relationships that are essential for the corresponding concerns. History-based aspects, as those proposed by the ASCOLA project-team provide a starting point that is not subject to this problem.

From a point of view of language design and implementation, aspect languages, as well as domain specific languages for distributed and concurrent environments share many characteristics with existing distributed languages: for instance, event monitoring is fundamental for pointcut matching, different synchronization strategies and strategies for code mobility [57] may be used in actions triggered by pointcuts. However, these relationships have only been explored to a small degree. Similarly, the formal semantics and formal properties of aspect languages have not been studied yet for the distributed case and only rudimentarily for the concurrent one [41], [55].

3.5. Security

Security properties and policies over complex service-oriented and standalone applications become ever more important in the context of asynchronous and decentralized communicating systems. Furthermore, they constitute prime examples of crosscutting functionalities that can only be modularized in highly insufficient ways with existing programming language and service models. Security properties and related properties, such as accountability properties, are therefore very frequently awkward to express and difficult to analyze and enforce (provided they can be made explicit in the first place).

Two main issues in this space are particularly problematic from a compositional point of view. First, information flow properties of programming languages, such as flow properties of Javascript [42], and service-based systems [46] are typically specially-tailored to specific properties, as well as difficult to express and analyze. Second, the enforcement of security properties and security policies, especially accountability-related properties [72], [78], is only supported using ad hoc means with rudimentary support for property verification.

The ASCOLA team has recently started to work on providing formal methods, language support and implementation techniques for the modular definition and implementation of information flow properties as well as policy enforcement in service-oriented systems as well as, mostly object-oriented, programming languages.

3.6. Capacity Planning for Large Scale Distributed System

Since the last decade, cloud computing has emerged as both a new economic model for software (provision) and as flexible tools for the management of computing capacity. Nowadays, the major cloud features have become part of the mainstream (virtualization, storage and software image management) and the big market players offer effective cloud-based solutions for resource pooling. It is now possible to deploy virtual infrastructures that involve virtual machines (VMs), middleware, applications, and networks in such a simple manner that a new problem has emerged over the last two years: VM sprawl (virtual machine proliferation) that consumes valuable computing, memory, storage and energy resources, thus menacing serious resource shortages. Scientific approaches that address VM sprawl are both based on classical administration techniques like the lifecycle management of a large number of VMs as well as the arbitration and the careful management of all resources consumed and provided by the hosting infrastructure (energy, power, computing, memory, network etc.).

The ASCOLA team investigates fundamental techniques for cloud computing and capacity planning, from infrastructures to the application level. Capacity planning is the process of planning for, analyzing, sizing, managing and optimizing capacity to satisfy demand in a timely manner and at a reasonable cost. Applied to distributed systems like clouds, a capacity planning solution must mainly provide the minimal set of resources necessary for the proper execution of the applications (i.e., to ensure service level agreement, SLA). The main challenges in this context are: scalability, fault tolerance and reactivity of the solution in a large-scale distributed system, the analysis and optimization of resources to minimize the cost (mainly costs related to the energy consumption of datacenters), as well as the profiling and adaptation of applications to ensure useful levels of quality of service (throughput, response time, availability etc.).

Our solutions are mainly based on virtualized infrastructures that we apply from the IaaS to the SaaS levels. We are mainly concerned by the management and the execution of the applications by harnessing virtualization capabilities, the investigation of alternative solutions that aim at optimizing the trade-off between performance and energy costs of both applications and cloud resources, as well as arbitration policies in the cloud in the presence of energy-constrained resources.

ATLANMOD Project-Team

3. Research Program

3.1. MDE Foundations

Traditionally, models were often used as initial design sketches mainly aimed for communicating ideas among developers. On the contrary, MDE promotes models as the primary artifacts that drive all software engineering activities (i.e. not only software development but also evolution, reverse engineering, interoperability and so on) and are considered as the unifying concept [41]. Therefore, rigorous techniques for model definition and manipulation are the basis of any MDE framework.

The MDE community distinguishes three levels of models: (terminal) model, metamodel, and metametamodel. A terminal model is a (partial) representation of a system/domain that captures some of its characteristics (different models can provide different knowledge views on the domain and be combined later on to provide a global view). In MDE we are interested in terminal models expressed in precise modeling languages. The abstract syntax of a language, when expressed itself as a model, is called a metamodel. A complete language definition is given by an abstract syntax (a metamodel), one or more concrete syntaxes (the graphical or textual syntaxes that designers use to express models in that language) plus one or more definition of its semantics. The relation between a model expressed in a language and the metamodel of that language is called *conformsTo*. Metamodels are in turn expressed in a modeling language called metamodeling language. Similar to the model/metamodel relationship, the abstract syntax of a metamodeling language is called a metametamodel and metamodels defined using a given metamodeling language must conform to its metametamodel. Terminal models, metamodels, and metametamodel form a three-level architecture with levels respectively named M1, M2, and M3. A formal definition of these concepts is provided in [49] and [42]. MDE promotes *unification by models*, like object technology proposed in the eighties *unification by objects* [39]. These MDE principles may be implemented in several standards. For example, OMG proposes a standard metametamodel called Meta Object Facility (MOF) while the most popular example of metamodel in the context of OMG standards is the UML metamodel.

In our view the main way to automate MDE is by providing model manipulation facilities in the form of model transformation operations that taking one or more models as input generate one or more models as output (where input and output models are not necessarily conforming to the same metamodel). More specifically, a model transformation Mt defines the production of a model Mb from a model Ma. When the source and target metamodels (MMs) are identical ($MMa = MMb$), we say that the transformation is endogenous. When this is not the case ($MMa \neq MMb$) we say the transformation is exogenous. An example of an endogenous transformation is a UML refactoring that transforms public class attributes into private attributes while adding accessor methods for each transformed attribute. Many other operations may be considered as transformations as well. For example verifications or measurements on a model can be expressed as transformations [44]. One can see then why large libraries of reusable modeling artifacts (mainly metamodels and transformations) will be needed.

Another important idea is the fact that a model transformation is itself a model [40]. This means that the transformation program Mt can be expressed as a model and as such conforms to a metamodel MMt. This allows an homogeneous treatment of all kinds of terminal models, including transformations. Mt can be manipulated using the same existing MDE techniques already developed for other kinds of models. For instance, it is possible to apply a model transformation Mt' to manipulate Mt models. In that case, we say that Mt' is a higher order transformation (HOT), i.e. a transformation taking other transformations (expressed as transformation models) as input or/and producing other transformations as output.

As MDE developed, it became apparent that this was a branch of language engineering [43]. In particular, MDE offers an improved way to develop DSLs (Domain-Specific Languages). DSLs are programming or modeling languages that are tailored to solve specific kinds of problems in contrast with General Purpose Languages (GPLs) that aim to handle any kind of problem. Java is an example of a programming GPL and UML an example of a modeling GPL. DSLs are already widely used for certain kinds of programming; probably the best-known example is SQL, a language specifically designed for the manipulation of relational data in databases. The main benefit of DSLs is that they allow everybody to write programs/models using the concepts that actually make sense to their domain or to the problem they are trying to solve (for instance Matlab has matrices and lets the user express operations on them, Excel has cells, relations between cells, and formulas and allows the expression of simple computations in a visual declarative style, etc.). As well as making domain code programmers more productive, DSLs also tend to offer greater optimization opportunities. Programs written with these DSLs may be independent of the specific hardware they will eventually run on. Similar benefits are obtained when using modeling DSLs. In MDE, new DSLs can be easily specified by using the metamodel concept to define their abstract syntax. Models specified with those DSLs can then be manipulated by means of model transformations (with ATL for example [48]).

When following the previously described principles, one may take advantage of the uniformity of the MDE organization. As an example, considering similarly models of the static architecture and models of the dynamic behavior of a system allows at the same time economy of concepts and economy of implementation.

The following sections describe the main MDE research challenges the team is addressing. They go beyond the development of core MDE techniques (topic on which the team, as mentioned above, has largely contributed in the past, and that we believe is quite well-covered already) and focus on new aspects that are critical for the successful application of MDE in industrial contexts.

3.2. Reverse Engineering

One important domain that is being investigated by the AtlanMod team is the reverse engineering of existing IT systems. We do believe that efficiently dealing with such legacy systems is one of the main challenges in Software Engineering and related industry today. Having a better understanding of these systems in order to document, maintain, improve or migrate them is thus a key requirement for both academic and industrial actors in this area. However, it is not an easy task and it still raises interesting challenging issues to be explored [46].

We have shown how reverse engineering practices may be advantageously revisited with the help of the MDE approach and techniques, applying (as base principle) the systematic representation as models of the required information discovered from the legacy software artifacts (e.g. source code, configuration files, documentation, metadata, etc). The rise in abstraction allowed by MDE can bring new hopes that reverse engineering is now able to move beyond more traditional ad-hoc practices. For instance, a industrial PhD in partnership with IBM France aimed to investigate the possibilities of conceptualizing a generic framework enabling the extraction of business rules from a legacy application, as much as possible, independently of the language used to code it. Moreover, different pragmatic solutions for improving the overall scalability when dealing with large-scale legacy systems (handling huge data volumes) are intensively studied by the team.

In this context, AtlandMod has set up within the past years and is still developing the open source Eclipse MoDisco project (see 5.2). MoDisco is notably being referenced by the OMG ADM (Architecture Driven Modernization) normalization task force as the reference implementation for several of its standard metamodels. It is also used practically and improved in various collaborative projects the team is currently involved in (e.g. FP7 ARTIST). Complementary to the work based on MoDisco, we have also been experimenting (still in an industrial context, cf. TEAP FUI project) on the related problem of data federation from heterogeneous sources in the domain of Enterprise Architecture. This has notably resulted in a prototype called EMF Views that can be practically used in such reverse engineering scenarios.

Reverse engineering techniques have also been used in the context of the Web. In the last years the development of Web APIs has become a discipline that companies have to master to succeed in the Web. The so-called API

economy requires, on the one hand, companies to provide access to their data by means of Web APIs and, on the other hand, web developers to study and integrate such APIs into their applications. The exchange of data with these APIs is usually performed by using JSON, a schemaless data format easy for computers to parse and use. While JSON data is easy to read, its structure is implicit, thus entailing serious problems when integrating APIs coming from different vendors. Web developers have therefore to understand the domain behind each API and study how they can be composed. We tackle this problem by developing a MDE-based process able to reverse engineer the domain of Web APIs and to identify composition links among them. The approach therefore allows developers to easily visualize what is behind the API and the connections points that may be used in their applications.

We have recently opened a new research line in the context software analysis, in particular, in the Open-Source Software (OSS) field. The development of OSS follows a collaborative model where any developer can contribute to the advance of the project. To enable this collaboration, OSS projects use a plethora of tools such as forums, issue-trackers and Q&A websites, that developers can adopt to coordinate each other in the development process. Such a collaboration environment includes adapted solutions and provides effective communication means, but also causes scattering of the collaboration data, which hamper the understanding of the whole development process (e.g., who is leading the development or making the decisions). In this context, we propose to use reverse engineering techniques to better understand how OSS projects are developed in a broad sense, thus taking into account the different collaboration tools used and how they influence in the development of OSS projects.

3.3. Security Engineering

Several components are required to build up a system security architecture, such as firewalls, database user access control, intrusion detection systems, and VPN (Virtual Private Network) routers. These components must be properly configured to provide an appropriate degree of security to the system. The configuration process is highly complex and error-prone. In most organizations, security components are either manually configured based on security administrators expertise and flair; or simply recycled from existing configurations already deployed in other systems (even if they may not be appropriated for the current one). These practices put at risk the security of the whole organization.

As a first step we intend to apply model-driven techniques for the extraction of high level model representations of security policies enforced by system components like networks of firewalls, RDBMS and CMSs. Firewalls, core components in network security systems, are generally configured by using very low level vendor specific rule-based languages, difficult to understand and to maintain. As a consequence, as the configuration files grow, understanding which security policy is being actually enforced or checking if inconsistencies has been introduced becomes a very complex and time consuming task. Similarly, in RDBMSs and CMSs policies are configured and stored by using different, often low-level, mechanisms.

We propose to raise the level of abstraction so that the user can deal directly with the high level policies. Once a model representation of the enforced policy is available, model-driven techniques will ease some of the tasks we need to perform, like consistency checking, validation, querying and visualization. Easy migration between different vendors will be also enabled.

As a further step we intend to apply model-driven techniques for the integration of the diverse security policies extracted from concrete system components. In the case of complex systems composed of a number of interacting heterogeneous subsystems, access-control is pervasive with respect to their architecture. As mentioned above, we can find access-control enforcement rules in different components placed at different architectural levels where rules in a component may impact the execution of the security rules of another component. In addition, the access-control techniques implemented in each component may follow different AC models in order to best suit the needs of the component. Thus, ideally, a global representation of the access-control policy of the whole system should be available, as analysing a component policy in isolation does not provide enough information. Unfortunately, most times this global policy is not explicit or is outdated. This step requires to unveil the implicit dependencies between the set of policies working in an encompassing

system, so that a model representing the global AC policy can be built and the global analysis of the AC security is enabled

3.4. Software Quality

As with any type of production, an essential part of software production is determining the quality of the software. The level of quality associated to a software product is inevitably tied to properties such as how well it was developed and how useful it is to its users. AtlanMod team focus on researching techniques for the formal verification and testing of software models and model transformations.

These techniques must be applied at the model level (to evaluate the quality of specific software designs) and at the metamodel level (to evaluate the quality of modeling languages). In both cases, the Object Constraint Language (OCL) of the OMG is widely accepted as a standard textual language to complement (meta)model specifications with all those rules/constraints that cannot be easily defined using graphical modeling constructs.

Among all possible properties to verify, we take as the basic property the *satisfiability* property, from which many others may be derived (as liveliness, redundancy, subsumption,...). Satisfiability checks whether it is possible to create a valid instantiation (i.e. one that respects all modeling constraints) of a give (meta)model. Satisfiability is an undedicable problem when general OCL constraints are used as part of the model definition.

To deal with this problem, the team maintains the tool EMFtoCSP which translates the model verification challenge into the domain of constraint logic programming (CLP) for which sophisticated decision procedures exist. The tool integrates the described functionality in the Eclipse Modeling Framework (EMF) and the Eclipse Modeling Tools (MDT), making the functionality available for MDE in practice.

To complement these formal verification techniques we are also working on testing techniques, specially to optimize the testing of model transformations. White-box testing for model transformations is a technique that involves the extraction of knowledge embedded in the transformation code to generate test models. In our work, we apply static analysis techniques to model transformation specifications and represent the extracted knowledge as partial models that can drive the generation of highly effective test models (specially in terms of coverage).

3.5. Collaborative Development

Software development processes are collaborative in nature. The active participation of end-users in the early phases of the software development life-cycle is key when developing software. Among other benefits, the collaboration promotes a continual validation of the software to be build, thus guaranteeing that the final software will satisfy the users' needs. In this context, we have opened two novel research lines focused on the collaborative development *in* MDE and the collaborative development *with* MDE. The former is aimed at promoting the collaboration in the context of MDE while the latter uses MDE techniques to promote the participation in software development processes.

Collaboration is important in the context of MDE, in particular, when creating Domain-Specific Modeling Languages (DSMLs) which are (modeling) languages specifically designed to carry out the tasks of a particular domain. While end-users are actually the experts of the domain for which a DSML is developed, their participation in the DSML specification process is still rather limited nowadays (they are normally only involved in providing domain knowledge or testing the resulting language). This means that the MDE technical experts and not end-users are the ones in control of the DSML construction and evolution. This is a problem because errors in understanding the domain may hamper the development process and the quality of the resulting DSML. Thus, it would be beneficial to promote a more active participation of end-users in the DSML development process.

We have been working on the required support to make effective this participation, in particular, we have developed Collaboro, an approach which enables the involvement of the community (i.e., end-users and developers) in the DSML creation process. Collaboro allows modeling the collaborations between community members taking place during the definition of a new DSML and supports both the collaborative definition of

the abstract (i.e., metamodel) and concrete (i.e., notation) syntaxes for DSMLs by providing specific constructs to enable the discussion. Thus, each community member will have the chance to request changes, propose solutions and give an opinion (and vote) about those from others. We believe this discussion will enrich the language definition significantly and ensure that the end result satisfies as much as possible the expectations of the end-users. Collaboro has also been extended to support the example-driven development of DSMLs, thus promoting the engagement of end-users in the process.

The lessons learnt from this MDE-focused collaboration research are now being applied to the more general context of software development. In particular, our interest is to study how software development processes are governed (i.e. how the collaboration among developers and user takes place). Any software development project has to cope with a huge number of tasks consisting of either implementing new issues or fixing bugs. Thus, effective and precise prioritization of these tasks is key for the success of the project. Governance rules enable the coordination of developers in order to advance the project. Despite their importance, in practice governance rules are hardly ever explicitly defined, specially in the context of Open Source Systems (OSS), where it is hard to find a explicit system-level design, a project plan, schedule or list of deliverables. To alleviate this situation, mechanisms to facilitate the communication and the assignment of work are considered crucial for the success of the development. Tracking and issue-tracking systems, mailing lists and forums are broadly used to manage the tasks to be performed. While these tools provide a convenient compartmentalization of work and effective means of communication, they fall short in providing adequate support for specifying and enforcing governance rules (e.g. supporting the voting of tasks, easy tracking of decisions made in the project, etc.).

Thus, we believe the explicit definition of governance rules along with the corresponding infrastructure to help developers follow them would have several benefits, including improvements in the transparency of the decision-making process, traceability (being able to track why a decision was made and who decided it) and the automation of the governance process (e.g. liberating developers from having to be aware and follow the rules manually, minimizing the risk of inconsistent behaviour in the evolution of the project). We resort on MDE techniques to tackle this problem and provide a DSL specially adapted to the domain of governance in software projects to let project managers easily define the governance rules of their projects.

3.6. Scalability

As MDE is increasingly applied to larger and more complex industrial applications, the current generation of modelling and model management technologies are being stressed to their limits in terms of their capacity to accommodate collaborative development, efficient management and persistence of models larger than a few hundreds of megabytes in size. Additional research and development is imperative in order to enable MDE to remain relevant with industrial practice and to continue delivering its widely recognised productivity, quality, and maintainability benefits. Achieving scalability in modelling and MDE involves being able to construct large models and domain-specific languages in a systematic manner, enabling teams of modellers to construct and refine large models in a collaborative manner, advancing the state-of-the-art in model querying and transformations tools so that they can cope with large models (of the scale of millions of model elements), and providing an infrastructure for efficient storage, indexing and retrieval of large models. AtlanMod wants to provide a solution for these aspects of scalability in MDE by extending the Eclipse modeling framework, to create an open-source solution to scalable modeling in industry.

3.7. Industrialization of open source tools

Research labs, as a source of innovation, are potential key actors of the Software Engineering market. However, an important collaborative effort with the other players in the software industry is still needed in order to actually transfer the corresponding techniques or technologies from the research lab to a company. Based on the AtlanMod concrete experience with the previously mentioned open source tools/projects, we have extracted a pragmatic approach [4] for transforming the results of scientific experimentation into practical industrial solutions.

While dealing with innovation, this approach is also innovation-driven itself, as the action is actually conducted by the research lab via a technology transfer. Three different partners are directly involved in this process, using open source as the medium for maintaining a constant interaction between all of them:

- **Use Case Provider.** Usually a company big enough to have to face real complex industrial scenarios which need to be solved (at least partially) by applying new innovative principles and techniques;
- **Research Lab.** Usually a group from a research institute (public or private) or university evaluating the scientific relevance of the problems, identifying the research challenges and prototyping possible solutions;
- **Technology Provider.** Usually a small or medium company, with a particular technical expertise on the given domain or Software Engineering field, building and delivering the industrial version of the designed solutions;

From our past and current experience, three main characteristics of this industrialization *business model* can be highlighted:

- **Win-win situation.** Each partner can actually focus on its core activity while also directly benefiting from the results obtained by the others (notably the research lab can continue to do research);
- **Application-driven context.** The end-user need is at the origin of the process, which finally makes the developed solution actually relevant;
- **Iterative process.** The fact of having three distinct partners requires different regular and consecutive exchanges between all of them.

AVALON Project-Team

3. Research Program

3.1. Energy Application Profiling and Modelization

International roadmaps schedule to build exascale systems by the 2018 time frame. According to the Top500 list published in November 2013, the most powerful supercomputer is the Tianhe-2 platform, a machine with more than 3,000,000 cores. It consumes more than 17 MW for a maximum performance of 33 PFlops while the Defense Advanced Research Projects Agency (DARPA) has set to 20 MW the maximum energy consumption of an exascale supercomputer [32].

Energy efficiency is therefore a major challenge for building next generation large scale platforms. The targeted platforms will gather hundreds of million cores, low power servers, or CPUs. Besides being very important, their power consumption will be dynamic and irregular.

Thus, to consume energy efficiently, we aim at investigating two research directions. First, we need to improve the measure, the understanding, and the analysis of the large-scale platform energy consumption. Unlike approaches [34] that mix the usage of internal and external wattmeters on a small set of resources, we target high frequency and precise internal and external energy measurements of each physical and virtual resources on large scale distributed systems.

Secondly, we need to find new mechanisms that consume less and better on such platforms. Combined with hardware optimizations, several works based on shutdown or slowdown approaches aim at reducing energy consumption of distributed platforms and applications. To consume less, we first plan to explore the provision of accurate estimation of the energy consumed by applications without pre-executing and knowing them while most of the works try to do it based on in-depth application knowledge (code instrumentation [37], phase detection for specific HPC applications [42], etc.). As a second step, we aim at designing a framework model that allows interactions, dialogues and decisions taken in cooperation between the user/application, the administrator, the resource manager, and the energy supplier. While smart grid is one of the last killer scenarios for networks, electrical provisioning of next generation large IT infrastructures remains a challenge.

3.2. Data-intensive Application Profiling, Modeling, and Management

Recently, the term “Big Data” has emerged to design data sets or collections so large that they become intractable for classical tools. This term is most of the time implicitly linked to “analytics” to refer to issues such as curation, storage, search, sharing, analysis, and visualization. However, the Big Data challenge is not limited to data-analytics, a field that is well covered by programming languages and run-time systems such as Map-Reduce. It also encompasses data-intensive applications. These applications can be sorted into two categories. In High Performance Computing (HPC), data-intensive applications leverage post-petascale infrastructures to perform highly parallel computations on large amount of data, while in High Throughput Computing (HTC), a large amount of independent and sequential computations are performed on huge data collections.

These two types of data-intensive applications (HTC and HPC) raise challenges related to profiling and modeling that the Avalon team proposes to address. While the characteristics of data-intensive applications are very different, our work will remain coherent and focused. Indeed, a common goal will be to acquire a better understanding of both the applications and the underlying infrastructures running them to propose the best match between application requirements and infrastructure capacities. To achieve this objective, we will extensively rely on logging and profiling in order to design sound, accurate, and validated models. Then, the proposed models will be integrated and consolidated within a single simulation framework (SimGrid). This will allow us to explore various potential “what-if?” scenarios and offer objective indicators to select interesting infrastructure configurations that match application specificities.

Another challenge is the ability to mix several heterogeneous infrastructure that scientists have at their disposal (*e.g.,* Grids, Clouds, and Desktop Grids) to execute data-intensive applications. Leveraging the aforementioned results, we will design strategies for efficient data management service for hybrid computing infrastructures.

3.3. Resourc-Agnostic Application Description Model

When programming in parallel, users expect to obtain performance improvement, whatever the cost is. For long, parallel machines have been simple enough to let a user program them given a minimal abstraction of their hardware. For example, MPI [36] exposes the number of nodes but hides the complexity of network topology behind a set of collective operations; OpenMP [40] simplifies the management of threads on top of a shared memory machine while OpenACC [39] aims at simplifying the use of GPGPU.

However, machines and applications are getting more and more complex so that the cost of manually handling an application is becoming very high [35]. Hardware complexity also stems from the unclear path towards next generations of hardware coming from the frequency wall: multi-core CPU, many-core CPU, GPGPUs, deep memory hierarchy, etc. have a strong impact on parallel algorithms. Hence, even though an abstract enough parallel language (UPC, Fortress, X10, etc.) succeeds, it will still face the challenge of supporting distinct codes corresponding to different algorithms corresponding to distinct hardware capacities.

Therefore, the challenge we aim to address is to define a model, for describing the structure of parallel and distributed applications that enables code variations but also efficient executions on parallel and distributed infrastructures. Indeed, this issue appears for HPC applications but also for cloud oriented applications. The challenge is to adapt an application to user constraints such as performance, energy, security, etc.

Our approach is to consider component based models [43] as they offer the ability to manipulate the software architecture of an application. To achieve our goal, we consider a "compilation" approach that transforms a resource-agnostic application description into a resource-specific description. The challenge is thus to determine a component based model that enables to efficiently compute application mapping while being tractable. In particular, it has to provide an efficient support with respect to application and resource elasticity, energy consumption and data management.

3.4. Application Mapping and Scheduling

This research axis is at the crossroad of the Avalon team. In particular, it gathers results of the three others research axis. We plan to consider application mapping and scheduling through the following three issues.

3.4.1. Application Mapping and Software Deployment

Application mapping and software deployment consist in the process of assigning distributed pieces of software to a set of resources. Resources can be selected according to different criteria such as performance, cost, energy consumption, security management, etc. A first issue is to select resources at application launch time. With the wide adoption of elastic platforms, *i.e.,* platforms that let the number of resources allocated to an application to be increased or decreased during its execution, the issue is also to handle resource selection at runtime.

The challenge in this context corresponds to the mapping of applications onto distributed resources. It will consist in designing algorithms that in particular take into consideration application profiling, modeling, and description.

A particular facet of this challenge is propose scheduling algorithms for dynamic and elastic platforms. As the amount of elements can vary, some kind of control of the platforms must be used accordingly to the scheduling.

3.4.2. Non-Deterministic Workflow Scheduling

Many scientific applications are described through workflow structures. Due to the increasing level of parallelism offered by modern computing infrastructures, workflow applications now have to be composed not only of sequential programs, but also of parallel ones. New applications are now built upon workflows with conditionals and loops (also called non-deterministic workflows).

These workflows can not be scheduled beforehand. Moreover cloud platforms bring on-demand resource provisioning and pay-as-you-go billing models. Therefore, there is a problem of resource allocation for non-deterministic workflows under budget constraints and using such an elastic management of resources.

Another important issue is data management. We need to schedule the data movements and replications while taking job scheduling into account. If possible, data management and job scheduling should be done at the same time in a closely coupled interaction.

3.4.3. Security Management in Cloud Infrastructure

Security has been proven to be sometimes difficult to obtain [41] and several issues have been raised in Clouds. Nowadays virtualization is used as the sole mechanism to secure different users sharing resources on Clouds. But, due to improper virtualization of all the components of Clouds (such as micro-architectural components), data leak and modification can occur. Accordingly, next-generation protection mechanisms are required to enforce security on Clouds and provide a way to cope with the current limitation of virtualization mechanisms.

As we are dealing with parallel and distributed applications, security mechanisms must be able to cope with multiple machines. Our approach is to combine a set of existing and novel security mechanisms that are spread in the different layers and components of Clouds in order to provide an in-depth and end-to-end security on Clouds. To do it, our first challenge is to define a generic model to express security policies.

Our second challenge is to work on security-aware resource allocation algorithms. The goal of such algorithms is to find a good trade-off between security and unshared resources. Consequently, they can limit resources sharing to increase security. It leads to complex trade-off between infrastructure consolidation, performance, and security.

CIDRE Project-Team

3. Research Program

3.1. Our perspective

For many aspects of our everyday life, we rely heavily on information systems, many of which are based on massively networked devices that support a population of interacting and cooperating entities. While these information systems become increasingly open and complex, accidental and intentional failures get considerably more frequent and severe.

Two research communities traditionally address the concern of accidental and intentional failures: the distributed computing community and the security community. While both these communities are interested in the construction of systems that are correct and secure, an ideological gap and a lack of communication exist between them that is often explained by the incompatibility of the assumptions each of them traditionally makes. Furthermore, in terms of objectives, the distributed computing community has favored systems availability while the security community has focused on integrity and confidentiality, and more recently on privacy.

By contrast with this traditional conception, we are convinced that by looking at information systems as a combination of possibly revisited basic protocols, each one specified by a set of properties such as synchronization and agreement, security properties should emerge. This vision is shared by others and in particular by Myers *et al.* [63], whose objectives are to explore new methods for constructing distributed systems that are trustworthy in the aggregate even when some nodes in the system have been compromised by malicious attackers.

In accordance with this vision, the first main characteristic of the CIDRE group is to gather researchers from the two aforementioned communities, in order to address intentional failures, using foundations and approaches coming from both communities.

The second main characteristic of the CIDRE group lies in the scope of the systems it considers. Indeed, we consider three complementary levels of study:

- The Node Level: The term node either refers to a device that hosts a network client or service or to the process that runs this client or service. Node security management must be the focus of a particular attention, since from the user point of view, security of his own devices is crucial. Sensitive information and services must therefore be locally protected against various forms of attacks. This protection may take a dual form, namely prevention and detection.
- The Group Level: Distributed applications often rely on the identification of sets of interacting entities. These subsets are either called groups, clusters, collections, neighborhoods, spheres, or communities according to the criteria that define the membership. Among others, the adopted criteria may reflect the fact that its members are administrated by a unique person, or that they share the same security policy. It can also be related to the localization of the physical entities, or the fact that they need to be strongly synchronized, or even that they share mutual interests. Due to the vast number of possible contexts and terminologies, we refer to a single type of set of entities, that we call set of nodes. We assume that a node can locally and independently identify a set of nodes and modify the composition of this set at any time. The node that manages one set has to know the identity of each of its members and should be able to communicate directly with them without relying on a third party. Despite these two restrictions, this definition remains general enough to include as particular cases most of the examples mentioned above. Of course, more restrictive behaviors can be specified by adding other constraints. We are convinced that security can benefit from the existence and the identification of sets of nodes of limited size as they can help in improving the efficiency of the detection and prevention mechanisms.

- The Open Network Level: In the context of large-scale distributed and dynamic systems, interaction with unknown entities becomes an unavoidable habit despite the induced risk. For instance, consider a mobile user that connects his laptop to a public Wifi access point to interact with his company. At this point, data (regardless if it is valuable or not) is updated and managed through non trusted undedicated entities (i.e., communication infrastructure and nodes) that provide multiple services to multiple parties during that user connection. In the same way, the same device (e.g., laptop, PDA, USB key) is often used for both professional and private activities, each activity accessing and manipulating decisive data.

The third characteristic of the CIDRE group is to focus on three different aspects of security, namely trust, intrusion detection, and privacy as well as on the bridges that exist between these aspects. Indeed, we believe that to study new security solutions for nodes, set of nodes and open network levels, one must take into account that it is now a necessity to interact with devices whose owners are unknown. To reduce the risk of relying on dishonest entities, a trust mechanism is an essential prevention tool that aims at measuring the capacity of a remote node to provide a service compliant with its specification. Such a mechanism should allow to overcome ill-founded suspicions and to be aware of established misbehaviors. To identify such misbehaviors, intrusion detection systems are necessary. Such systems aim at detecting, by analyzing data flows, whether violations of the security policies have occurred. Finally, Privacy, which is now recognized as a fundamental individual right, should be respected despite the presence of tools and systems that continuously observe or even control users actions or behaviors.

3.2. Intrusion Detection

By exploiting vulnerabilities in operating systems, applications, or network services, an attacker can defeat preventive security mechanisms and violate the security policy of the whole system. The goal of intrusion detection systems (IDS) is to detect, by analyzing some data generated on a monitored system, violations of the security policy. From our point of view, while useful in practice, misuse detection is intrinsically limited. Indeed, it requires to update the signatures database in real-time similarly to what has to be done for antivirus tools. Given that there are thousands of machines that are every day victims of malware, such an approach may appear as insufficient especially due to the incredible expansion of malware, drastically limiting the capabilities of human intervention and response. The CIDRE group takes the alternative approach, namely the anomaly approach, which consists in detecting a deviation from a referenced behavior. Specifically, we propose to study three complementary methods:

- Illegal Flow Detection: This first method intends to detect information flows that violate the security policy [66], [62]. Our goal is here to detect information flows in the monitored system that are allowed by the access control mechanism, but are illegal from the security policy point of view.
- Data Corruption Detection: This second method aims at detecting intrusions that target specific applications, and make them execute illegal actions by using these applications incorrectly [60], [65]. This approach complements the previous one in the sense that the incorrect use of the application can possibly be legal from the point of view of the information flows and access control mechanisms, but is incorrect considering the security policy.
- Visualization: This third method relies on the capacity of human beings in detecting patterns and outliers in datasets when these datasets are properly visually represented. Human beings also know pieces of contextual information that are very difficult to formalize so as to make them usable by a computer. Visualization is therefore a very useful complementary tool to detect abnormal events in real time (monitoring), to search for malicious events in log files (data exploration and forensics) and to communicate results (reporting).

In these approaches, the access control mechanisms or the monitored applications can be either configured and executed on a single node, or distributed on a set of nodes. Thus, our approach must be studied at least at these two levels.

Here are some concrete examples of our research objectives (both short term and long term objectives) in the intrusion detection field:

- At node level, we apply the defensive programming approach (coming from the dependability field) to data corruption detection. The challenge is to determine which invariant/properties must be and can be verified either at runtime or statically. Regarding illegal flow detection, we try to extend this method to build anti-viruses by determining viruses signatures.
- At the set of nodes level, we revisit the distributed problems such as clock synchronization, logical clocks, consensus, properties detection, to extend the solutions proposed at node levels to cope with distributed flow control checking mechanisms. Regarding illegal flow detection, we study the collaboration and consistency at the node and set of nodes levels to obtain a global intrusion detection mechanism. Regarding the data corruption detection approach, our challenge is to identify local predicates/properties/invariants so that global predicates/properties/invariants would emerge at the system level.

3.3. Privacy

In our world of ubiquitous technologies, each individual constantly leaves digital traces related to his activities and interests which can be linked to his identity. The protection of privacy is one of the greatest challenge that lies ahead and also an important condition for the development of the Information Society. Moreover, due to legality and confidentiality issues, issues linked to privacy emerge naturally for applications working on sensitive data, such as medical records of patients or proprietary datasets of enterprises. Privacy Enhancing Technologies (PETs) are generally designed to respect both the principles of data minimization and data sovereignty. The data minimization principle states that only the information necessary to complete a particular application should be disclosed (and no more). This principle is a direct application of the legitimacy criteria defined by the European data protection directive (Article 7). This directive is currently being revised into a regulation that is going to strengthen the privacy rights of individuals and puts forward the concept of "privacy-by-design", which integrates the privacy aspects into the conception phase of a service or product. The data sovereignty principle states that data related to an individual belong to him and that he should stay in control of how this data is used and for which purpose. This principle can be seen as an extension of many national legislations on medical data that consider that a patient record belongs to the patient, and not to the doctors that create or update it, nor to the hospital that stores it. A fundamental hindrance to the achievement of sovereignty is that the trust assumptions given to external entities are often too optimistic, and thus they are many realistic situations in which they might betrayed.

In the CIDRE project, we investigate PETs operating at three different levels (node, set of nodes or open distributed system) and that are generally based on a mix of different foundations such as cryptographic techniques, security policies and access control mechanisms just to name a few. Examples of domains in which privacy and utility aspects collide and that are studied within the context of CIDRE include: identity management, location-based services, social networks, distributed systems and data mining. Here are some concrete examples of our research goals in the privacy field:

- at the node level, we design privacy-preserving identification scheme, automated reasoning on privacy policies [64], and policy-based adaptive PETs.
- at the set of nodes level, we augment distributed algorithms with privacy properties such as anonymity, unlinkability and unobservability.
- at the open distributed system level, we target both privacy concerns linked to disclosure of location (that typically occur in location-based services) and privacy issues in social networks. In the former case, we adopt a sanitization approach while in the latter one we consider privacy policies at user level, and their enforcement by all the intervening actors (e.g, at the level of the social network providers, of intermediate servers or of individual peers). We design novel algorithms for the resolution of privacy policy conflicts between autonomous entities, taking new concepts into consideration, such as the notion of equity in the context of an access control decision.

3.4. Trust Management

While the distributed computing community relies on the trustworthiness of its algorithms to ensure systems availability, the security community historically makes the hypothesis of a Trusted Computing Base (TCB) that contains the security mechanisms (such as access controls, and cryptography) implementing the security policy. Unfortunately, as information systems get increasingly complex and open, the TCB management may itself get very complex, dynamic and error-prone. From our point of view, an appealing approach is to distribute and manage the TCB on each node and to leverage the trustworthiness of the distributed algorithms to strengthen each node's TCB. Accordingly, the CIDRE group studies automated trust management systems at all the three identified levels:

- at the node level, such a system should allow each node to evaluate by itself the trustworthiness of its neighborhood and to self-configure the security mechanisms it implements;
- at the group level, such a system might rely on existing trust relations with other nodes of the group to enhance the significance and the reliability of the gathered information;
- at the open network level, such a system should rely on reputation mechanisms to estimate the trustworthiness of the peers the node interacts with. The system might also benefit from the information provided by *a priori* trusted peers that, for instance, would belong to the same group (see previous item).

For the last two items, the automated trust management system will de facto follow the distributed computing approach. As such, emphasis will be put on the trustworthiness of the designed distributed algorithms. Thus, the proposed approach will provide both the adequate security mechanisms and a trustworthy distributed way of managing them. Regarding trust management, we still have research goals that are to be tackled. We briefly list hereafter some of our short and long term objectives at node, group and open networks levels:

1. At node level, we investigate how implicit trust relationships identified and deduced by a node during its interactions with its neighborhood could be explicitly used by the node (for instance by means of a series of rules) to locally evaluate the trustworthiness of its neighborhood. The impact of trust on the local security policy, and on its enforcement will be studied accordingly.
2. At the set of nodes level, we take advantage of the pre-existing trust relationship among the set of nodes to design composition mechanisms that would guarantee that automatically configured security policies are consistent with each group member security policy.
3. At the open distributed system level, we design reputation mechanisms to both defend the system against specific attacks (whitewashing, bad mouthing, ballot stuffing, isolation) by relying on the properties guaranteed at nodes and set of nodes levels, and guaranteeing persistent and safe feedback, and for specific cases in guaranteeing the right to be forgotten (i.e., the right to data erasure).

COAST Team

3. Research Program

3.1. Introduction

Our scientific foundations are grounded on distributed collaborative systems supported by sophisticated data sharing mechanisms and on service oriented computing with an emphasis on orchestration and on non functional properties.

Distributed collaborative systems enable distributed group work supported by computer technologies. Designing such systems require an expertise in Distributed Systems and in Computer-supported collaborative activities research area. Besides theoretical and technical aspects of distributed systems, design of distributed collaborative systems must take into account the human factor to offer solutions suitable for users and groups. The COAST team vision is to move away from a centralized authority based collaboration towards a decentralized collaboration where users have full control over their data that they can store locally and decide with whom to share them. The Coast team investigates the issues related to the management of distributed shared data and coordination between users and groups.

Service oriented Computing [27] is an established domain on which the ECOO, SCORE and now the Coast team have been contributing for a long time. It refers to the general discipline that studies the development of computer applications on the web. A service is an independent software program with a specific functional context and capabilities published as a service contract (or more traditionally an API). A service composition aggregates a set of services and coordinates their interactions. The scale, the autonomy of services, the heterogeneity and some design principles underlying Service Oriented Computing open new research questions that are at the basis of our research. They span the disciplines of distributed computing, software engineering and computer supported collaborative work (CSCW). Our approach to contribute to the general vision of Service Oriented Computing and more generally to the emerging discipline of Service Science has been and is still to focus on the question of the efficient and flexible construction of reliable and secure high level services through the coordination/orchestration/composition of other services provided by distributed organizations or people.

3.2. Consistency Models for Distributed Collaborative Systems

Collaborative systems are distributed systems that allow users to share data. One important issue is to manage consistency of shared data according to concurrent access. Traditional consistency criteria such as locking, serializability, linearizability are not adequate for collaborative systems.

Causality, Convergence and Intention preservation (CCI) [30] are more suitable for developing middleware for collaborative applications.

We develop algorithms for ensuring CCI properties on collaborative distributed systems. Constraints on the algorithms are different according to the type of distributed system and type of data. The distributed system can be centralized, decentralized or peer-to-peer. The type of data can include strings, growable arrays, ordered trees, semantic graphs and multimedia data.

3.3. Optimistic Replication

Replication of data among different nodes of a network allows improving reliability, fault-tolerance, and availability. When data are mutable, consistency among the different replicas must be ensured. Pessimistic replication is based on the principle of single-copy consistency while optimistic replication allows the replicas to diverge during a short time period. The consistency model for optimistic replication [29] is called eventual consistency, meaning that replicas are guaranteed to converge to the same value when the system is idle.

Our research focuses on the two most promising families of optimistic replication algorithms for ensuring CCI:

- the operational transformation (OT) algorithms [25]
- the algorithms based on commutative replicated data types (CRDT) [28].

Operational transformation algorithms are based on the application of a transformation function when a remote modification is integrated into the local document. Integration algorithms are generic, being parametrized by operational transformation functions which depend on replicated document types. The advantage of these algorithms is their genericity. These algorithms can be applied to any data type and they can merge heterogeneous data in a uniform manner.

Commutative replicated data types is a new class of algorithms initiated by WOOT [26] a first algorithm designed WithOut Operational Transformations. They ensure consistency of highly dynamic content on peer-to-peer networks. Unlike traditional optimistic replication algorithms, they can ensure consistency without concurrency control. CRDT algorithms rely on natively commutative operations defined on abstract data types such as lists or ordered trees. Thus, they do not require a merge algorithm or an integration procedure.

3.4. Process Orchestration and Management

Process Orchestration and Management is considered as a core discipline behind Service Management and Computing. It includes the analysis, the modelling, the execution, the monitoring and the continuous improvement of enterprise processes and is for us a central domain of studies.

Much efforts has been devoted in the past years to establish standard business process models founded on well grounded theories (e.g. Petri Nets) that meet the needs of both business analysts but also of software engineers and software integrators. This has lead to heated debate as both points of view are very difficult to reconciliate between the analyst side and the IT side. On one side, the business people in general require models that are easy to use and understand and that can be quickly adapted to exceptional situations. On the other side, IT people need models with an operational semantic in order to be able transform them into executable artefacts. Part of our work has been an attempt to reconcile these point of views. It has lead to the development of Bonita product and more recently on our work in crisis management where the same people are designing, executing and monitoring the process as it executes. But more generally, and at a larger scale, we have been considering the problem of process spanning the barriers of organisations. This leads us to consider the more general problem of service composition as a way to coordinate inter organisational construction of applications providing value based on the composition of lower level services [24].

3.5. Service Composition

More and more, we are considering processes as pieces of software whose execution traverse the boundaries of organisations. This is especially true with service oriented computing where processes compose services produced by many organisations. We tackle this problem from very different perspectives, trying to find the best compromise between the need for privacy of internal processes from organisations and the necessity to publicize large part of them, proposing to distribute the execution and the orchestration of processes among the organisations themselves, and attempting to ensure non functional properties in this distributed setting [23].

Non functional aspects of service composition relate to all the properties and service agreements that one want to ensure and that are orthogonal to the actual business but that are important when a service is selected and integrated in a composition. This includes transactional context, security, privacy, and quality of service in general. Defining and orchestrating services on a large scale while providing the stakeholders with some strong guarantees on their execution is a first class problem for us. For a long time, we have proposed models and solutions to ensure that some properties (e.g. transactional properties) were guaranteed on process execution, either through design or through the definition of some protocols. Our work has also been extended to the problems of security, privacy and service level agreement among partners. These questions are still central in our work. Then, one major problem of current approaches is to monitor the execution

of the compositions, integrating the distributed dimension. This problem can be tackled using event-based algorithms and techniques. Using our event oriented composition framework DISC, we have obtained new results dedicated to the runtime verification of violations in service choreographies.

COATI Project-Team

3. Research Program

3.1. Research Program

Members of COATI have a good expertise in the design and management of wired and wireless backbone, backhaul, broadband, and complex networks. On the one hand, we cope with specific problems such as energy efficiency in backhaul and backbone networks, routing reconfiguration in connection oriented networks (MPLS, WDM), traffic agregation in SONET networks, compact routing in large-scale networks, survivability to single and multiple failures, etc. These specific problems often come from questions of our industrial partners. On the other hand, we study fondamental problems mainly related to routing and reliability that appear in many networks (not restricted to our main fields of applications) and that have been widely studied in the past. However, previous solutions do not take into account the constraints of current networks/traffic such as their huge size and their dynamics. COATI thus puts a significant research effort in the following directions:

- **Energy efficiency** at both the design and management levels. More precisely, we plan to develop accurate modeling of the power consumption of various parts and components of the networks through measurement done in collaboration with industrial partners (Alcatel-Lucent, 3Roam, Orange labs, etc.). Then, we shall propose new designs of the networks and new routing algorithms in order to lower the power consumption.
- **Larger networks:** Another challenge one has to face is the increase in size of practical instances. It is already difficult, if not impossible, to solve practical instances optimally using existing tools. Therefore, we have to find new ways to solve problems using reduction and decomposition methods, characterization of polynomial instances (which are surprisingly often the practical ones), or algorithms with acceptable practical performances.
- **Stochastic behaviors:** Larger topologies mean frequent changes due to traffic and radio fluctuations, failures, maintenance operations, growth, routing policy changes, etc. We aim at including these stochastic behaviors in our combinatorial optimization process to handle the dynamics of the system and to obtain robust designs of networks.

CTRL-A Exploratory Action

3. Research Program

3.1. Modeling and control techniques for autonomic computing

3.1.1. Continuous control

Continuous control was used to control computer systems only very recently and in few occasions, despite the promising results that were obtained. This is probably due to many reasons, but the most important seems to be the difficulty by both communities to transform a computer system problem into an automatic control problem. The aim of the team is to explore how to formalize typical autonomic commuting cases into typical control problems. Many new methodological tools will probably be useful for that, e.g., we can cite the hybrid system approach, predictive control or event-based control approach. Computer systems are not usual for the control system community and they often present non-conventional control aspects like saturation control. New methodological tools are required for an efficient use of continuous-time control in computer science.

3.1.2. Discrete control

Discrete control techniques are explored at long-term, to integrate more control in the BZR language, and adress more general control issues, wider than BZR's limitations. Directions are : expressiveness (taking into account in the LTS models value domains of the variables in the program) ; adaptive control (where the controller itself can dynamically switch between differents modes) ; distributed control (for classes of problems where communicating controllers can be designed) ; optimal control (w.r.t. weight functions, on states, transitions, and paths, with multicriteria techniques) ; timed and hybrid control bringing a new dimension for modeling and control, giving solutions where discrete models fail.

3.2. Design and programming for autonomic computing

3.2.1. Reactive programming

Autonomic systems are intrinsically reconfigurable. To describe, specify or design these systems, there is a need to take into account this reconfigurability, within the programming languages used. We propose to consider the reconfigurability of systems from the angle of two properties: the notion of time, as we want to describe the state and behavior of the system before, and after its reconfiguration; the notion of dynamicity of the system, i.e., considering that the system's possible behaviors throughout execution are not completely known, neither at design-time nor at initial execution state. To describe and design such reactive systems, we propose to use the synchronous paradigm. It has been successfully used, in industry, for the design of embedded systems. It allows the description of behaviors based on a specific model of time (discrete time scale, synchronous parallel composition), providing properties which are important w.r.t. the safety of the described system: reactivity, determinism, preservation of safety properties by parallel composition (with other parts of the system or with its environment). Models and languages for control, proposed in this framework, provide designers, experts of the application domain, with a user-friendly access to highly technical formal methods of DCS, by encapsulating them in the compilation of concrete programming languages, generating concrete executable code. They are based on discrete models, but also support programming of sampled continuous controllers.

3.2.2. Component-based approach and domain-specific languages

For integration of the previous control kernels into wider frameworks of reconfigurable systems, they have to be integrated in a design flow, and connected on the one side with higher-level specification languages (with help of DSLs), and on the other side with the generated code level target execution machines. This calls for the adoption of a component-based approach with necessary features, available typically in Fractal, for explicitly identifying the control interfaces and mechanisms.

Structuring and instrumentation for controllability will involve encapsulation of computations into components, specification of their local control (activation, reconfiguration, suspension, termination), and exporting appropriate interfaces (including behavior abstraction). Modeling the configurations space requires determining the controlled aspects (e.g., heterogenous CPUs loads, fault-tolerance and variability, memory, energy/power consumption, communication/bandwidth, QoS level) and their control points, as well as APIs for monitors and actions. Compilation and execution will integrate this in a complete design flow involving : extraction of a reactive model from components; instrumentation of execution platforms to be controllable; combination with other controllers; general "glue" and wrapper code.

Integration of reactive languages and control techniques in component-based systems brings interesting questions of co-existence w.r.t. other approaches like Event-Condition-Action (ECA) rules, or Complex Event Processing (CPE).

3.3. Infrastructure-level support for autonomic computing

The above general kernel of model-based control techniques can be used in a range of different computing infrastructures, representing complementary targets and abstraction levels, exploring the two axes :

- from hardware, to operating system/virtual machine, to middleware, to applications/service level;
- across different criteria for adaptation: resources and energy, quality of service, dependability.

3.3.1. Software and adaptive systems

Autonomic administration loops at operating systems or middleware level are already very widespread. An open problem remains in design techniques for controllers with predictability and safety, e.g. w.r.t. the reachable states. We want to contribute to the topic of discrete control techniques for these systems, and tackle e.g. problems of coordination of multiple autonomic loops in data-centers, as in the ANR project CtrlGreen. Another target application is the control of clusters in map-reduce applications. The objective is to use continuous time control in order to tune finely the number of required clusters for an application running on a map-reduce server. This will use results of the ANR project MyCloud that enables to simulate clients on a real map-reduce server. On a longer term, we are interested in control problems in administration loops of event-based virtual machines, or in the deployment of massively parallel computation of the Cloud.

3.3.2. Hardware and reconfigurable architectures

Reconfigurable architectures based on Field Programmable Gate Arrays (FPGA) are an active research area, where infrastructures are more and more supportive of reconfiguration, but its correct control remains an important issue. Work has begun in the ANR Famous project on identifying domain-specific control criteria and objectives, monitors and management APIs, and on integrating control techniques in the high-level RecoMARTE environment. On a longer term, we want to work on methods and tools for the programming of **multicore architectures**, exploiting the reconfigurability potentials and issues (because of variability, loss of cores), e.g. in our cooperation with ST Microelectronics, using a Fractal-based programming framework in the P2012 project, and in cooperation with Inria Lille (Adam), or with the CEA and TIMA on integrating control loops in the architecture for a fine control of the energy and of the required nodes for running a given application task.

3.3.3. Applications and autonomic systems

In autonomic systems, control systems remain a lively source of inspiration, partly because the notion of control loop implementation is known and practiced naturally. On a wider scale, we started a cooperation with Orange Labs on "intelligent" building automation and control for the Smart Grid, through modeling and control of appliances w.r.t. their power consumption modes, at home, building, and city levels. Other partners on these topics are CEA LETI/DACLE and Schneider Electric.

We could explore more systems and applications e.g., Human-Machine Interfaces, or the orchestration of services. They can help design more general solutions, and result in a more complete methodology.

DANTE Team

3. Research Program

3.1. Graph-based signal processing

Participants: Christophe Crespelle, Éric Fleury, Paulo Gonçalves, Márton Karsai, Benjamin Girault.

Evolving networks can be regarded as "*out of equilibrium*" systems. Indeed, their dynamics is typically characterized by non standard and intricate statistical properties, such as non-stationarity, long range memory effects, intricate space and time correlations.

Analyzing, modeling, and even defining adapted concepts for dynamic graphs is at the heart of DANTE. This is a largely open question that has to be answered by keeping a balance between specificity (solutions triggered by specific data sets) and generality (universal approaches disconnected from social realities). We will tackle this challenge from a graph-based signal processing perspective involving signal analysts and computer scientists, together with experts of the data domain application. One can distinguish two different issues in this challenge, one related to the graph-based organisation of the data and the other to the time dependency that naturally exits in the dynamic graph object. In both cases, a number of contributions can be found in the literature, albeit in different contexts. In our application domain, high-dimensional data "naturally reside" on the vertices of weighted graphs. The emerging field of signal processing on graphs merges algebraic and spectral graph theoretic concepts with computational harmonic analysis to process such signals on graphs [48].

As for the first point, adapting well-founded signal processing techniques to data represented as graphs is an emerging, yet quickly developing field which has already received key contributions. Some of them are very general and delineate ambitious programs aimed at defining universal, generally unsupervised methods for exploring high-dimensional data sets and processing them. This is the case for instance of the « diffusion wavelets » and « diffusion maps » pushed forward at Yale and Duke [33]. Others are more traditionally connected with standard signal processing concepts, in the spirit of elaborating new methodologies via some bridging between networks and time series, see, *e.g.*, ([43] and references therein). Other viewpoints can be found as well, including multi-resolution Markov models [51], Bayesian networks or distributed processing over sensor networks [42]. Such approaches can be particularly successful for handling static graphs and unveiling aspects of their organisation in terms of dependencies between nodes, grouping, etc. Incorporating possible time dependencies within the whole picture calls however for the addition of an extra dimension to the problem "as it would be the case when switching from one image to a video sequence", a situation for which one can imagine to take advantage of the whole body of knowledge attached to non-stationary signal processing [34].

3.2. Theory and Structure of dynamic Networks

Participants: Christophe Crespelle, Éric Fleury, Anthony Busson, Márton Karsai.

Characterization of the dynamics of complex networks. We need to focus on intrinsic properties of evolving/dynamic complex networks. New notions (as opposed to classical static graph properties) have to be introduced: rate of vertices or links appearances or disappearances, the duration of link presences or absences. Moreover, more specific properties related to the dynamics have to be defined and are somehow related to the way to model a dynamic graph.

Through the systematic analysis and characterization of static network representations of many different systems, researchers of several disciplines have unveiled complex topologies and heterogeneous structures, with connectivity patterns statistically characterized by heavy-tails and large fluctuations, scale-free properties and non trivial correlations such as high clustering and hierarchical ordering [45]. A large amount of work has been devoted to the development of new tools for statistical characterisation and modelling of networks, in order to identify their most relevant properties, and to understand which growth mechanisms could lead to these properties. Most of those contributions have focused on static graphs or on dynamic process (*e.g.* diffusion) occurring on static graphs. This has called forth a major effort in developing the methodology to characterize the topology and temporal behavior of complex networks [45], [36], [52], [41], to describe the observed structural and temporal heterogeneities [30], [36], [31], to detect and measure emerging community structures [35], [49], [50], to see how the functionality of networks determines their evolving structure [40], and to determine what kinds of correlations play a role in their dynamics [37], [39], [44].

The challenge is now to extend this kind of statistical characterization to dynamical graphs. In other words, links in dynamic networks are temporal events, called contacts, which can be either punctual or last for some period of time. Because of the complexity of this analysis, the temporal dimension of the network is often ignored or only roughly considered. Therefore, fully taking into account the dynamics of the links into a network is a crucial and highly challenging issue.

Another powerful approach to model time-varying graphs is via activity driven network models. In this case, the only assumption relates to the distribution of activity rates of interacting entities. The activity rate is realistically broadly distributed and refers to the probability that an entity becomes active and creates a connection with another entity within a unit time step [47]. Even the generic model is already capable to recover some realistic features of the emerging graph, its main advantage is to provide a general framework to study various types of correlations present in real temporal networks. By synthesizing such correlations (*e.g.* memory effects, preferential attachment, triangular closing mechanisms, ...) from the real data, we are able to extend the general mechanism and build a temporal network model, which shows certain realistic feature in a controlled way. This can be used to study the effect of selected correlations on the evolution of the emerging structure [38] and its co-evolution with ongoing processes like spreading phenomena, synchronisation, evolution of consensus, random walk etc. [38], [46]. This approach allows also to develop control and immunisation strategies by fully considering the temporal nature of the backgrounding network.

3.3. Distributed Algorithms for dynamic networks: regulation, adaptation and interaction

Participants: Thomas Begin, Anthony Busson, Paulo Gonçalves, Isabelle Guérin Lassous.

Dedicated algorithms for dynamic networks. First, the dynamic network object itself trigger original algorithmic questions. It mainly concerns distributed algorithms that should be designed and deployed to efficiently measure the object itself and get an accurate view of its dynamic behavior. Such distributed measure should be "transparent", that is, it should introduce no bias or at least a bias that is controllable and corrigible. Such problem is encountered in all distributed metrology measures / distributed probes: P2P, sensor network, wireless network, QoS routing... This question raises naturally the intrinsic notion of adaptation and control of the dynamic network itself since it appears that autonomous networks and traffic aware routing are becoming crucial.

Communication networks are dynamic networks that potentially undergo high dynamicity. The dynamicity exhibited by these networks results from several factors including, for instance, changes in the topology and varying workload conditions. Although most implemented protocols and existing solutions in the literature can cope with a dynamic behavior, the evolution of their behavior operates identically whatever the actual properties of the dynamicity. For instance, parameters of the routing protocols (*e.g.* hello packets transmission frequency) or routing methods (*e.g.* reactive / proactive) are commonly hold constant regardless of the nodes mobility. Similarly, the algorithms ruling CSMA/CA (*e.g.* size of the contention window) are tuned identically and they do not change according to the actual workload and observed topology.

Dynamicity in computer networks tends to affect a large number of performance parameters (if not all) coming from various layers (viz. physical, link, routing and transport). To find out which ones matter the most for our intended purpose, we expect to rely on the tools developed by the two former axes. These quantities should capture and characterize the actual network dynamicity. Our goal is to take advantage of this latter information in order to refine existing protocols, or even to propose new solutions. More precisely, we will attempt to associate "fundamental" changes occurring in the underlying graph of a network (reported through graph-based signal tools) to quantitative performance that are matter of interests for networking applications and the end-users. We expect to rely on available testbeds such as Senslab and FIT to experiment our solutions and ultimately validate our approach.

DIANA Team

3. Research Program

3.1. Service Transparency

Transparency is to provide network users and application developers with reliable information about the current or predicted quality of their communication services, and about potential leakages of personal information, or of other information related to societal interests of the user as a "connected citizen" (e.g. possible violation of network neutrality, opinion manipulation). Service transparency therefore means to provide information meaningful to users and application developers, such as quality of experience, privacy leakages, or opinion manipulation, etc. rather than network-level metrics such as available bandwidth, loss rate, delay or jitter.

The Internet is built around a best effort routing service that does not provide any guarantee to end users in terms of quality of service (QoS). The simplicity of the Internet routing service is at the root of its huge success. Unfortunately, a simple service means unpredicted quality at the access. Even though a considerable effort is done by operators and content providers to optimise the Internet content delivery chain, mainly by over-provisioning and sophisticated engineering techniques, service degradation is still part of the Internet. The proliferation of wireless and mobile access technologies, and the versatile nature of Internet traffic, make end users quality of experience (QoE) forecast even harder. As a matter of fact, the Internet is missing a dedicated measurement plane that informs the end users on the quality they obtain and in case of substantial service degradation, on the origin of this degradation. The mPlane FP7 project (http://www.ict-mplane.eu) is devoted to building a distributed measurement infrastructure to perform active, passive and hybrid measurements in the wired Internet. However, the problem is exacerbated with modern terminals such as smartphones or tablets that do not facilitate the task for end users (they even make it harder) as they focus on simplifying the interface and limiting the control on the network, whereas the Internet behind is still the same in terms of the quality it provides. Interestingly, this same observation explains the existing difficulty to detect and prevent privacy leaks. We argue that the lack of transparency for diagnosing QoE and for detecting privacy leaks have the same root causes and can be solved using common primitives. For instance, in both cases, it is important to be able to link data packets to an application. Indeed, as the network can only access data packets, there must be a way to bind these packets to an application (to understand users QoE for this application or to associate a privacy leak to an application). This is however a complex task as the traffic might be obfuscated or encrypted. Our objectives in the research direction are the following:

- Design and develop measurement tools providing transparency, in spite of current complexity
- Deploy those measurement tools at the Internet's edge and make them useful for end users
- Propose measurements plane as an overlay or by exploiting in-network functionalities
- Adapt measurements techniques to network architectural change
- Provide measurements as native functionality in future network architecture

3.2. Open network architecture

We are surrounded by personal content of all types: photos, videos, documents, etc. The volume of such content is increasing at a fast rate, and at the same time, the spread of such content among all our connected devices (mobiles, storage devices, set-top boxes, etc) is also increasing. All this complicates the control of personal content by the user both in terms of access and sharing with other users. The access of the personal content in a seamless way independently of its location is a key challenge for the future of networks. Proprietary solutions exist, but apart from fully depending on one of them, there is no standard plane in the Internet for a seamless access to personal content. Therefore, providing network architectural support to design and develop content access and sharing mechanisms is crucial to allow users control their own data over heterogeneous underlying network or cloud services.

On the other hand, privacy is a growing concern for states, administrations, and companies. Indeed, for instance the French CNIL (entity in charge of citizens privacy in computer systems) puts privacy at the core of its activities by defining rules on any stored and collected private data. Also, companies start to use privacy preserving solutions as a competitive advantage. Therefore, understanding privacy leaks and preventing them is a problem that can already find support. However, all end-users do not *currently* put privacy as their first concern. Indeed, in face of two services with one of higher quality, they usually prefer the highest quality one whatever the privacy implication. This was, for instance, the case between the Web search service of Google that is more accurate but less privacy preserving than Bing. This is also the case for cloud services such as iCloud or Dropbox that are much more convenient than open source solutions, but very bad in terms of privacy. Therefore, to reach end-users, any privacy preserving solutions must offer a service equivalent to the best existing services.

We consider that it will be highly desirable for Internet users to be able to *easily* move their content from a provider to another and therefore not to depend on a content provider or a social network monopoly. This requires that the network provides built-in architectural support for content networking.

In this research direction, we will define a new *service abstraction layer* (SAL) that could become the new waist of the network architecture with network functionalities below (IP, SDN, cloud) and applications on top. SAL will define different services that are of use to all Internet users for accessing and sharing data (seamless content localisation and retrieval, privacy leakage protection, transparent vertical and horizontal handover, etc.). The biggest challenge here is to cope in the same time with large number of content applications requirements and high underlying networks heterogeneity while still providing efficient applications performance. This requires careful definition of the services primitives and the parameters to be exchanged through the service abstraction layer.

Two concurring factors make the concept behind SAL feasible and relevant today. First, the notion of scalable network virtualization that is a required feature to deploy SAL in real networks today has been discussed recently only. Second, the need for new services abstraction is recent. Indeed, fifteen years ago the Internet for the end-users was mostly the Web. Only eight years ago smartphones came into the picture of the Internet boosting the number of applications with new functionalities and risks. Since a few years, many discussions in the network communities took place around the actual complexity of the Internet and the difficulty to develop applications. Many different approaches have been discussed (such as CCN, SDN) that intend to solve only part of the complexity. SAL takes a broader architectural look at the problem and considers solutions such as CCN as mere use cases. Our objectives in this research direction include the following:

- Identify common key networking services required for content access and sharing
- Detect and prevent privacy leaks for content communication
- Enhance software defined networks for large scale heterogeneous environments
- Design and develop open Content Networking architecture
- Define a service abstraction layer as the thin waist for the future content network architecture
- Test and deploy different applications using SAL primitives on heterogeneous network technologies

3.3. Methodology

We follow an experimental approach that can be described in the following techniques:

- Measurements: the aim is to get a better view of a problem in quantifiable terms. Depending on the field of interest, this may involve large scale distributed systems crawling tools; active probing techniques to infer the status and properties of a complex and non controllable system as the Internet; or even crowdsourcing-based deployments for gathering data on real-users environments or behaviours.
- Experimental evaluation: once a new idea has been designed and implemented, it is of course very desirable to assess and quantify how effective it can be, before being able to deploy it on any realistic scale. This is why a wide range of techniques can be considered for getting early, yet as significant as possible, feedback on a given paradigm or implementation. The spectrum for such techniques span from simulations to real deployments in protected and/or controlled environments.

DIONYSOS Project-Team

3. Research Program

3.1. Introduction

The scientific foundations of our work are those of network design and network analysis. Specifically, this concerns the principles of packet switching and in particular of IP networks (protocol design, protocol testing, routing, scheduling techniques), and the mathematical and algorithmic aspects of the associated problems, on which our methods and tools are based.

These foundations are described in the following paragraphs. We begin by a subsection dedicated to Quality of Service (QoS) and Quality of Experience (QoE), since they can be seen as unifying concepts in our activities. Then we briefly describe the specific sub-area of model evaluation and about the particular multidisciplinary domain of network economics.

3.2. Quality of Service and Quality of Experience

Since it is difficult to develop as many communication solutions as possible applications, the scientific and technological communities aim towards providing general *services* allowing to give to each application or user a set of properties nowadays called "Quality of Service" (QoS), a terminology lacking a precise definition. This QoS concept takes different forms according to the type of communication service and the aspects which matter for a given application: for performance it comes through specific metrics (delays, jitter, throughput, etc.), for dependability it also comes through appropriate metrics: reliability, availability, or vulnerability, in the case for instance of WAN (Wide Area Network) topologies, etc.

QoS is at the heart of our research activities: We look for methods to obtain specific "levels" of QoS and for techniques to evaluate the associated metrics. Our ultimate goal is to provide tools (mathematical tools and/or algorithms, under appropriate software "containers" or not) allowing users and/or applications to attain specific levels of QoS, or to improve the provided QoS, if we think of a particular system, with an optimal use of the resources available. Obtaining a good QoS level is a very general objective. It leads to many different areas, depending on the systems, applications and specific goals being considered. Our team works on several of these areas. We also investigate the impact of network QoS on multimedia payloads to reduce the impact of congestion.

Some important aspects of the behavior of modern communication systems have subjective components: the quality of a video stream or an audio signal, *as perceived by the user*, is related to some of the previous mentioned parameters (packet loss, delays, ...) but in an extremely complex way. We are interested in analyzing these types of flows from this user-oriented point of view. We focus on the *user perceived quality*, the main component of what is nowadays called Quality of Experience (in short, QoE), to underline the fact that, in this case, we want to center the analysis on the user. In this context, we have a global project called PSQA, which stands for Pseudo-Subjective Quality Assessment, and which refers to a methodology allowing to automatically measure QoE.

Another special case to which we devote research efforts in the team is the analysis of qualitative properties related to interoperability assessment. This refers to the act of determining if end-to-end functionality between at least two communicating systems is as required by the base standards for those systems. Conformance is the act of determining to what extent a single component conforms to the individual requirements of the standard it is based on. Our purpose is to provide such a formal framework (methods, algorithms and tools) for interoperability assessment, in order to help in obtaining efficient interoperability test suites for new generation networks, mainly around IPv6-related protocols. The interoperability test suites generation is based on specifications (standards and/or RFCs) of network components and protocols to be tested.

3.3. Stochastic modeling

The scientific foundations of our modeling activities are composed of stochastic processes theory and, in particular, Markov processes, queuing theory, stochastic graphs theory, etc. The objectives are either to develop numerical solutions, or analytical ones, or possibly discrete event simulation or Monte Carlo (and Quasi-Monte Carlo) techniques. We are always interested in model evaluation techniques for dependability and performability analysis, both in static (network reliability) and dynamic contexts (depending on the fact that time plays an explicit role in the analysis or not). We look at systems from the classical so-called *call level*, leading to standard models (for instance, queues or networks of queues) and also at the *burst level*, leading to *fluid models*.

In recent years, our work on the design of the topologies of WANs led us to optimization techniques, in particular in the case of very large optimization problems, usually formulated in terms of graphs. The associated methods we are interested in are composed of simulated annealing, genetic algorithms, TABU search, etc. For the time being, we have obtained our best results with GRASP techniques.

Network pricing is a good example of a multi-disciplinary research activity half-way between applied mathematics, economy and networking, centered on stochastic modeling issues. Indeed, the Internet is facing a tremendous increase of its traffic volume. As a consequence, real users complain that large data transfers take too long, without any possibility to improve this by themselves (by paying more, for instance). A possible solution to cope with congestion is to increase the link capacities; however, many authors consider that this is not a viable solution as the network must respond to an increasing demand (and experience has shown that demand of bandwidth has always been ahead of supply), especially now that the Internet is becoming a commercial network. Furthermore, incentives for a fair utilization between customers are not included in the current Internet. For these reasons, it has been suggested that the current flat-rate fees, where customers pay a subscription and obtain an unlimited usage, should be replaced by usage-based fees. Besides, the future Internet will carry heterogeneous flows such as video, voice, email, web, file transfers and remote login among others. Each of these applications requires a different level of QoS: for example, video needs very small delays and packet losses, voice requires small delays but can afford some packet losses, email can afford delay (within a given bound) while file transfer needs a good average throughput and remote login requires small round-trip times. Some pricing incentives should exist so that each user does not always choose the best QoS for her application and so that the final result is a fair utilization of the bandwidth. On the other hand, we need to be aware of the trade-off between engineering efficiency and economic efficiency; for example, traffic measurements can help in improving the management of the network but is a costly option. These are some of the various aspects often present in the pricing problems we address in our work. More recently, we have switched to the more general field of network economics, dealing with the economic behavior of users, service providers and content providers, as well as their relations.

DIVERSE Project-Team

3. Research Program

3.1. Scientific background

3.1.1. Model-driven engineering

Model-Driven Engineering (MDE) aims at reducing the accidental complexity associated with developing complex software-intensive systems (e.g., use of abstractions of the problem space rather than abstractions of the solution space) [148]. It provides DIVERSE with solid foundations to specify, analyze and reason about the different forms of diversity that occur through the development lifecycle. A primary source of accidental complexity is the wide gap between the concepts used by domain experts and the low-level abstractions provided by general-purpose programming languages [116]. MDE approaches address this problem through modeling techniques that support separation of concerns and automated generation of major system artifacts from models (*e.g.,* test cases, implementations, deployment and configuration scripts). In MDE, a model describes an aspect of a system and is typically created or derived for specific development purposes [94]. Separation of concerns is supported through the use of different modeling languages, each providing constructs based on abstractions that are specific to an aspect of a system. MDE technologies also provide support for manipulating models, for example, support for querying, slicing, transforming, merging, and analyzing (including executing) models. Modeling languages are thus at the core of MDE, which participates to the development of a sound *Software Language Engineering*[0], including an unified typing theory that integrate models as first class entities [151].

Incorporating domain-specific concepts and high-quality development experience into MDE technologies can significantly improve developer productivity and system quality. Since the late nineties, this realization has led to work on MDE language workbenches that support the development of domain-specific modeling languages (DSMLs) and associated tools (*e.g.,* model editors and code generators). A DSML provides a bridge between the field in which domain experts work and the implementation (programming) field. Domains in which DSMLs have been developed and used include, among others, automotive, avionics, and the emerging cyber-physical systems. A study performed by Hutchinson et al. [123] provides some indications that DSMLs can pave the way for wider industrial adoption of MDE.

More recently, the emergence of new classes of systems that are complex and operate in heterogeneous and rapidly changing environments raises new challenges for the software engineering community. These systems must be adaptable, flexible, reconfigurable and, increasingly, self-managing. Such characteristics make systems more prone to failure when running and thus the development and study of appropriate mechanisms for continuous design and run-time validation and monitoring are needed. In the MDE community, research is focused primarily on using models at design, implementation, and deployment stages of development. This work has been highly productive, with several techniques now entering a commercialization phase. As software systems are becoming more and more dynamic, the use of model-driven techniques for validating and monitoring run-time behavior is extremely promising [131].

3.1.2. Variability modeling

While the basic vision underlying *Software Product Lines* (SPL) can probably be traced back to David Parnas seminal article [141] on the Design and Development of Program Families, it is only quite recently that SPLs are emerging as a paradigm shift towards modeling and developing software system families rather than individual systems [139]. SPL engineering embraces the ideas of mass customization and software reuse. It focuses on the means of efficiently producing and maintaining multiple related software products, exploiting what they have in common and managing what varies among them.

[0] See http://planet-sl.org

Several definitions of the *software product line* concept can be found in the research literature. Clements *et al.* define it as *a set of software-intensive systems sharing a common, managed set of features that satisfy the specific needs of a particular market segment or mission and are developed from a common set of core assets in a prescribed way* [138]. Bosch provides a different definition [103]: *A SPL consists of a product line architecture and a set of reusable components designed for incorporation into the product line architecture. In addition, the PL consists of the software products developed using the mentioned reusable assets*. In spite of the similarities, these definitions provide different perspectives of the concept: *market-driven*, as seen by Clements *et al.*, and *technology-oriented* for Bosch.

SPL engineering is a process focusing on capturing the *commonalities* (assumptions true for each family member) and *variability* (assumptions about how individual family members differ) between several software products [110]. Instead of describing a single software system, a SPL model describes a set of products in the same domain. This is accomplished by distinguishing between elements common to all SPL members, and those that may vary from one product to another. Reuse of core assets, which form the basis of the product line, is key to productivity and quality gains. These core assets extend beyond simple code reuse and may include the architecture, software components, domain models, requirements statements, documentation, test plans or test cases.

The SPL engineering process consists of two major steps:

1. **Domain Engineering**, or *development for reuse*, focuses on core assets development.
2. **Application Engineering**, or *development with reuse*, addresses the development of the final products using core assets and following customer requirements.

Central to both processes is the management of **variability** across the product line [118]. In common language use, the term *variability* refers to *the ability or the tendency to change*. Variability management is thus seen as the key feature that distinguishes SPL engineering from other software development approaches [104]. Variability management is thus growingly seen as the cornerstone of SPL development, covering the entire development life cycle, from requirements elicitation [153] to product derivation [158] to product testing [137], [136].

Halmans *et al.* [118] distinguish between *essential* and *technical* variability, especially at requirements level. Essential variability corresponds to the customer's viewpoint, defining what to implement, while technical variability relates to product family engineering, defining how to implement it. A classification based on the dimensions of variability is proposed by Pohl *et al.* [143]: beyond **variability in time** (existence of different versions of an artifact that are valid at different times) and **variability in space** (existence of an artifact in different shapes at the same time) Pohl *et al.* claim that variability is important to different stakeholders and thus has different levels of visibility: **external variability** is visible to the customers while **internal variability**, that of domain artifacts, is hidden from them. Other classification proposals come from Meekel *et al.* [129] (feature, hardware platform, performances and attributes variability) or Bass *et al.* [92] who discuss about variability at the architectural level.

Central to the modeling of variability is the notion of *feature*, originally defined by Kang *et al.* as: *a prominent or distinctive user-visible aspect, quality or characteristic of a software system or systems* [125]. Based on this notion of *feature*, they proposed to use a *feature model* to model the variability in a SPL. A feature model consists of a *feature diagram* and other associated information: *constraints* and *dependency rules*. Feature diagrams provide a *graphical tree-like notation depicting the hierarchical organization of high level product functionalities* represented as features. The root of the tree refers to the complete system and is progressively decomposed into more refined features (tree nodes). Relations between nodes (features) are materialized by *decomposition edges* and *textual constraints*. Variability can be expressed in several ways. Presence or absence of a feature from a product is modeled using *mandatory* or *optional features*. Features are graphically represented as rectangles while some graphical elements (e.g., unfilled circle) are used to describe the variability (e.g., a feature may be optional).

Features can be organized into *feature groups*. Boolean operators *exclusive alternative (XOR)*, *inclusive alternative (OR)* or *inclusive (AND)* are used to select one, several or all the features from a feature group.

Dependencies between features can be modeled using *textual constraints*: *requires* (presence of a feature requires the presence of another), *mutex* (presence of a feature automatically excludes another). Feature attributes can be also used for modeling quantitative (e.g., numerical) information. Constraints over attributes and features can be specified as well.

Modeling variability allows an organization to capture and select which version of which variant of any particular aspect is wanted in the system [104]. To implement it cheaply, quickly and safely, redoing by hand the tedious weaving of every aspect is not an option: some form of automation is needed to leverage the modeling of variability [96], [112]. Model Driven Engineering (MDE) makes it possible to automate this weaving process [124]. This requires that models are no longer informal, and that the weaving process is itself described as a program (which is as a matter of facts an executable meta-model [133]) manipulating these models to produce for instance a detailed design that can ultimately be transformed to code, or to test suites [142], or other software artifacts.

3.1.3. Component-based software development

Component-based software development [152] aims at providing reliable software architectures with a low cost of design. Components are now used routinely in many domains of software system designs: distributed systems, user interaction, product lines, embedded systems, etc. With respect to more traditional software artifacts (e.g., object oriented architectures), modern component models have the following distinctive features [111]: description of requirements on services required from the other components; indirect connections between components thanks to ports and connectors constructs [127]; hierarchical definition of components (assemblies of components can define new component types); connectors supporting various communication semantics [107]; quantitative properties on the services [101].

In recent years component-based architectures have evolved from static designs to dynamic, adaptive designs (e.g., SOFA [107], Palladio [97], Frascati [134]). Processes for building a system using a statically designed architecture are made of the following sequential lifecycle stages: requirements, modeling, implementation, packaging, deployment, system launch, system execution, system shutdown and system removal. If for any reason after design time architectural changes are needed after system launch (e.g., because requirements changed, or the implementation platform has evolved, etc) then the design process must be reexecuted from scratch (unless the changes are limited to parameter adjustment in the components deployed).

Dynamic designs allow for *on the fly* redesign of a component based system. A process for dynamic adaptation is able to reapply the design phases while the system is up and running, without stopping it (this is different from stop/redeploy/start). This kind of process supports *chosen adaptation*, when changes are planned and realized to maintain a good fit between the needs that the system must support and the way it supports them [126]. Dynamic component-based designs rely on a component meta-model that supports complex life cycles for components, connectors, service specification, etc. Advanced dynamic designs can also take platform changes into account at run-time, without human intervention, by adapting themselves [109], [155]. Platform changes and more generally environmental changes trigger *imposed adaptation*, when the system can no longer use its design to provide the services it must support. In order to support an eternal system [99], dynamic component based systems must separate architectural design and platform compatibility. This requires support for heterogeneity, since platform evolutions can be partial.

The Models@runtime paradigm denotes a model-driven approach aiming at taming the complexity of dynamic software systems. It basically pushes the idea of reflection one step further by considering the reflection layer as a real model "something simpler, safer or cheaper than reality to avoid the complexity, danger and irreversibility of reality [146]". In practice, component-based (and/or service-based) platforms offer reflection APIs that make it possible to introspect the system (which components and bindings are currently in place in the system) and dynamic adaptation (by applying CRUD operations on these components and bindings). While some of these platforms offer rollback mechanisms to recover after an erroneous adaptation, the idea of Models@runtime is to prevent the system from actually enacting an erroneous adaptation. In other words, the "model at run-time" is a reflection model that can be uncoupled (for reasoning, validation, simulation purposes) and automatically resynchronized.

Heterogeneity is a key challenge for modern component based system. Until recently, component based techniques were designed to address a specific domain, such as embedded software for command and control, or distributed Web based service oriented architectures. The emergence of the Internet of Things paradigm calls for a unified approach in component based design techniques. By implementing an efficient separation of concern between platform independent architecture management and platform dependent implementations, *Models@runtime* is now established as a key technique to support dynamic component based designs. It provides DIVERSE with an essential foundation to explore an adaptation envelop at run-time.

Search Based Software Engineering [120] has been applied to various software engineering problems in order to support software developers in their daily work. The goal is to automatically explore a set of alternatives and assess their relevance with respect to the considered problem. These techniques have been applied to craft software architecture exhibiting high quality of services properties [117]. Multi Objectives Search based techniques [114] deal with optimization problem containing several (possibly conflicting) dimensions to optimize. These techniques provide DIVERSE with the scientific foundations for reasoning and efficiently exploring an envelope of software configurations at run-time.

3.1.4. Validation and verification

Validation and verification (V&V) theories and techniques provide the means to assess the validity of a software system with respect to a specific correctness envelop. As such, they form an essential element of DIVERSE's scientific background. In particular, we focus on model-based V&V in order to leverage the different models that specify the envelop at different moments of the software development lifecycle.

Model-based testing consists in analyzing a formal model of a system (*e.g.*, activity diagrams, which capture high-level requirements about the system, statecharts, which capture the expected behavior of a software module, or a feature model, which describes all possible variants of the system) in order to generate test cases that will be executed against the system. Model-based testing [154] mainly relies on model analysis, constraint solving [113] and search-based reasoning [128]. DIVERSE leverages in particular the applications of model-based testing in the context of highly-configurable systems and [156] interactive systems [130] as well as recent advances based on diversity for test cases selection [121].

Nowadays, it is possible to simulate various kinds of models. Existing tools range from industrial tools such as Simulink, Rhapsody or Telelogic to academic approaches like Omega [140], or Xholon [0]. All these simulation environments operate on homogeneous environment models. However, to handle diversity in software systems, we also leverage recent advances in heterogeneous simulation. Ptolemy [106] proposes a common abstract syntax, which represents the description of the model structure. These elements can be decorated using different directors that reflect the application of a specific model of computation on the model element. Metropolis [93] provides modeling elements amenable to semantically equivalent mathematical models. Metropolis offers a precise semantics flexible enough to support different models of computation. ModHel'X [119] studies the composition of multi-paradigm models relying on different models of computation.

Model-based testing and simulation are complemented by runtime fault-tolerance through the automatic generation of software variants that can run in parallel, to tackle the open nature of software-intensive systems. The foundations in this case are the seminal work about N-version programming [91], recovery blocks [144] and code randomization [95], which demonstrated the central role of diversity in software to ensure runtime resilience of complex systems. Such techniques rely on truly diverse software solutions in order to provide systems with the ability to react to events, which could not be predicted at design time and checked through testing or simulation.

3.1.5. Empirical software engineering

The rigorous, scientific evaluation of DIVERSE's contributions is an essential aspect of our research methodology. In addition to theoretical validation through formal analysis or complexity estimation, we also aim at applying state-of-the-art methodologies and principles of empirical software engineering. This approach encompasses a set of techniques for the sound validation contributions in the field of software engineering,

[0] http://www.primordion.com/Xholon/

ranging from statistically sound comparisons of techniques and large-scale data analysis to interviews and systematic literature reviews [149], [147]. Such methods have been used for example to understand the impact of new software development paradigms [105]. Experimental design and statistical tests represent another major aspect of empirical software engineering. Addressing large-scale software engineering problems often requires the application of heuristics, and it is important to understand their effects through sound statistical analyses [90].

3.2. Research axis

Figure 1 illustrates the four dimensions of software diversity, which form the core research axis of DIVERSE: the **diversity of languages** used by the stakeholders involved in the construction of these systems; the **diversity of features** required by the different customers; the **diversity of runtime environments** in which software has to run and adapt; the **diversity of implementations** that are necessary for resilience through redundancy. These four axis share and leverage the scientific and technological results developed in the area of model-driven engineering in the last decade. This means that all our research activities are founded on sound abstractions to reason about specific aspects of software systems, compose different perspectives and automatically generate parts of the system.

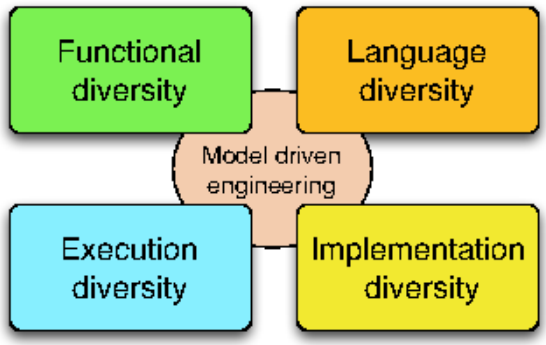


Figure 1. The four research axis of DIVERSE, which rely on a MDE scientific background

3.2.1. Software Language Engineering

The engineering of systems involves many different stakeholders, each with their own domain of expertise. Hence more and more organizations are adopting Domain Specific Modeling Languages (DSMLs) to allow domain experts to express solutions directly in terms of relevant domain concepts [148], [116]. This new trend raises new challenges about designing DSMLs, evolving a set of DSMLs and coordinating the use of multiple DSLs for both DSL designers and DSL users.

3.2.1.1. Challenges

Reusability of software artifacts is a central notion that has been thoroughly studied and used by both academics and industrials since the early days of software construction. Essentially, designing reusable artifacts allows the construction of large systems from smaller parts that have been separately developed and validated, thus reducing the development costs by capitalizing on previous engineering efforts. However, it is still hardly possible for language designers to design typical language artifacts (e.g. language constructs, grammars, editors or compilers) in a reusable way. The current state of the practice usually prevents the reusability of language artifacts from one language to another, consequently hindering the emergence of real engineering techniques around software languages. Conversely, concepts and mechanisms that enable artifacts reusability abound in the software engineering community.

Variability concerns in modeling languages occur in the definition of the abstract and concrete syntax as well as in the specification of the language's semantics. The major challenges met when addressing the need for variability are: (i) set principles for modeling language units that support the modular specification of a modeling language; and (ii) design mechanisms to assemble these units in a complete language, according to the set of authorized variation points for the modeling language family.

A new generation of complex software-intensive systems (for example smart health support, smart grid, building energy management, and intelligent transportation systems) presents new opportunities for leveraging modeling languages. The development of these systems requires expertise in diverse domains. Consequently, different types of stakeholders (e.g., scientists, engineers and end-users) must work in a coordinated manner on various aspects of the system across multiple development phases. DSMLs can be used to support the work of domain experts who focus on a specific system aspect, but they can also provide the means for coordinating work across teams specializing in different aspects and across development phases. The support and integration of DSMLs leads to what we call **the globalization of modeling languages**, *i.e.* the use of multiple languages for the coordinated development of diverse aspects of a system. One can make an analogy with world globalization in which relationships are established between sovereign countries to regulate interactions (e.g., travel and commerce related interactions) while preserving each country's independent existence.

3.2.1.2. Scientific objectives

We address reuse and variability challenges through the investigation of the time-honored concepts of substitutability, inheritance and components, evaluate their relevance for language designers and provide tools and methods for their inclusion in software language engineering. We will develop novel techniques for the modular construction of language extensions with the support of model syntactical variability. From the semantics perspective, we investigate extension mechanisms for the specification of variability in operational semantics, focusing on static introduction and heterogeneous models of computation. The definition of variation points for the three aspects of the language definition provides the foundations for the novel concept Language Unit (LU) as well as suitable mechanisms to compose such units.

We explore the necessary breakthrough in software languages to support modeling and simulation of heterogeneous and open systems. This work relies on the specification of executable domain specific modeling languages (DSMLs) to formalize the various concerns of a software-intensive system, and of models of computation (MoCs) to explicitly model the concurrency, time and communication of such DSMLs. We develop a framework that integrates the necessary foundations and facilities for designing and implementing executable and concurrent domain-specific modeling languages. It also provides unique features to specify composition operators between (possibly heterogeneous) DSMLs. Such specifications are amenable to support the edition, execution, graphical animation and analysis of heterogeneous models. The objective is to provide both a significant improvement of MoCs and DSMLs design and implementation; and the simulation based validation and verification of complex systems.

We see an opportunity for the automatic diversification of programs' computation semantics, for example through the diversification of compilers or virtual machines. The main impact of this artificial diversity is to provide flexible computation and thus ease adaptation to different execution conditions. A combination of static and dynamic analysis could support the identification of what we call *plastic computation zones* in the code. We identify different categories of such zones: (i) areas in the code in which the order of computation can vary (e.g., the order in which a block of sequential statements is executed); (ii) areas that can be removed, keeping the essential functionality [150] (e.g., skip some loop iterations); (iii) areas that can replaced by alternative code (e.g., replace a try-catch by a return statement). Once we know which zones in the code can be randomized, it is necessary to modify the model of computation to leverage the computation plasticity. This consists in introducing variation points in the interpreter to reflect the diversity of models of computation. Then, the choice of a given variation is performed randomly at run-time.

3.2.2. Variability Modeling and Engineering

The systematic modeling of variability in software systems has emerged as an effective approach to document and reason about software evolutions and heterogeneity (*cf.* Section 3.1.2). Variability modeling character-

izes an "envelope" of possible software variations. The industrial use of variability models and their relation to software artifact models require a complete engineering framework, including composition, decomposition, analysis, configuration and artifact derivation, refactoring, re-engineering, extraction, and testing. This framework can be used both to tame imposed diversity and to manage chosen diversity.

3.2.2.1. Challenges

A fundamental problem is that the **number of variants** can be exponential in the number of options (features). Already with 300 boolean configuration options, approximately 10^{90} configurations exist – more than estimated count of atoms in the universe. Domains like automotive or operating systems have to manage more than 10000 options (e.g., Linux). Practitioners face the challenge of developing billions of variants. It is easy to forget a necessary constraint, leading to the synthesis of unsafe variants, or to under-approximate the capabilities of the software platform. Scalable modelling techniques are therefore crucial to specify and reason about a very large set of variants.

Model-driven development supports two ways to deal with the increasing number of concerns in complex systems: (1) multi-view modeling, *i.e.* when modeling each concern separately, and variability modeling. However, there is little support to combine both approaches consistently. Techniques to integrate both approaches will enable the construction of a consistent set of views and variation points in each view.

The design, construction and maintenance of software families have a major impact on **software testing**. Among the existing challenges, we can cite: the selection of test cases for a specific variant; the evolution of test suites with integration of new variants; the combinatorial explosion of the number of software configurations to be tested. Novel model-based techniques for test generation and test management in a software product line context are needed to overcome state-of-the-art limits we already observed in some projects.

3.2.2.2. Scientific objectives

We aim at developing scalable techniques to automatically analyze variability models and their interactions with other views on the software intensive system (requirements, architecture, design). These techniques provide two major advancements in the state of the art: (1) an extension of the semantics of variability models in order to enable the definition of attributes (*e.g.*, cost, quality of service, effort) on features and to include these attributes in the reasoning; (2) an assessment of the consistent specification of variability models with respect to system views (since variability is orthogonal to system modeling, it is currently possible to specify the different models in ways that are semantically meaningless). The former aspect of analysis is tackled through constraint solving and finite-domain constraint programming, while the latter aspect is investigated through automatic search-based techniques (similar to genetic algorithms) for the exploration of the space of interaction between variability and view models.

We aim to develop procedures to reverse engineer dependencies and features' sets from existing software artefacts – be it source code, configuration files, spreadsheets (e.g., product comparison matrices) or requirements. We expect to scale up (e.g., for extracting a very large number of variation points) and guarantee some properties (e.g., soundness of configuration semantics, understandability of ontological semantics). For instance, when building complex software-intensive systems, textual requirements are captured in very large quantities of documents. In this context, adequate models to formalize the organization of requirements documents and automated techniques to support impact analysis (in case of changes in the requirements) have to be developed.

We aim at developing sound methods and tools to integrate variability management in model-based testing activities. In particular, we will leverage requirement models as an essential asset to establish formal relations between variation points and test models. These relations will form the basis for novel algorithms that drive the systematic selection of test configurations that satisfy well-defined test adequacy criteria as well as the generation of test cases for a specific product in the product line.

3.2.3. Heterogeneous and dynamic software architectures

Flexible yet dependable systems have to cope with heterogeneous hardware execution platforms ranging from smart sensors to huge computation infrastructures and data centers. Evolutions range from a mere change in the system configuration to a major architectural redesign, for instance to support addition of new features

or a change in the platform architecture (new hardware is made available, a running system switches to low bandwidth wireless communication, a computation node battery is running low, etc). In this context, we need to devise formalisms to reason about the impact of an evolution and about the transition from one configuration to another. It must be noted that this axis focuses on the use of models to drive the evolution from design time to run-time. Models will be used to (i) systematically define predictable configurations and variation points through which the system will evolve; (ii) develop behaviors necessary to handle unpredicted evolutions.

3.2.3.1. Challenges

The main challenge is to provide new homogeneous architectural modelling languages and efficient techniques that enable continuous software reconfiguration to react to changes. This work handles the challenges of handling the diversity of runtime infrastructures and managing the cooperation between different stakeholders. More specifically, the research developed in this axis targets the following dimensions of software diversity.

Platform architectural heterogeneity induces a first dimension of imposed diversity (type diversity). Platform reconfigurations driven by changing resources define another dimension of diversity (deployment diversity). To deal with these imposed diversity problems, we will rely on model based runtime support for adaptation, in the spirit of the dynamic distributed component framework developed by the Triskell team. Since the runtime environment composed of distributed, resource constrained hardware nodes cannot afford the overhead of traditional runtime adaptation techniques, we investigate the design of novel solutions relying on models@runtime and on specialized tiny virtual machines to offer resource provisioning and dynamic reconfigurations. In the next two years this research will be supported by the InfraJVM project.

Diversity can also be an asset to optimize software architecture. Architecture models must integrate multiple concerns in order to properly manage the deployment of software components over a physical platform. However, these concerns can contradict each other (*e.g.*, accuracy and energy). In this context, we investigate automatic solutions to explore the set of possible architecture models and to establish valid trade-offs between all concerns in case of changes.

3.2.3.2. Scientific objectives

Automatic synthesis of optimal software architectures. Implementing a service over a distributed platform (*e.g.*, a pervasive system or a cloud platform) consists in deploying multiple software components over distributed computation nodes. We aim at designing search-based solutions to (i) assist the software architect in establishing a good initial architecture (that balances between different factors such as cost of the nodes, latency, fault tolerance) and to automatically update the architecture when the environment or the system itself change. The choice of search-based techniques is motivated by the very large number of possible software deployment architectures that can be investigated and that all provide different trade-offs between qualitative factors. Another essential aspect that is supported by multi-objective search is to explore different architectural solutions that are not necessarily comparable. This is important when the qualitative factors are orthogonal to each other, such as security and usability for example.

Flexible software architecture for testing and data management. As the number of platforms on which software runs increases and different software versions coexist, the demand for testing environments also increases. For example, to test a software patch or upgrade, the number of testing environments is the product of the number of running environments the software supports and the number of coexisting versions of the software. Based on our first experiment on the synthesis of cloud environment using architectural models, our objective is to define a set of domain specific languages to catch the requirement and to design cloud environments for testing and data management of future internet systems from data centers to things. These languages will be interpreted to support dynamic synthesis and reconfiguration of a testing environment.

Runtime support for heterogeneous environments. Execution environments must provide a way to account or reserve resources for applications. However, current execution environments such as the Java Virtual Machine do not clearly define a notion of application: each framework has its own definition. For example, in OSGi, an application is a component, in JEE, an application is most of the time associated to a class loader, in the Multi-Tasking Virtual machine, an application is a process. The challenge consists in defining an execution environment that provides direct control over resources (CPU, Memory, Network I/O) independently from the

definition of an application. We propose to define abstract resource containers to account and reserve resources on a distributed network of heterogeneous devices.

3.2.4. Diverse implementations for resilience

Open software-intensive systems have to evolve over their lifetime in response to changes in their environment. Yet, most verification techniques assume a closed environment or the ability to predict all changes. Dynamic changes and evolutions thus represent a major challenge for these techniques that aim at assessing the correctness and robustness of the system. On the one hand, DIVERSE will adapt V&V techniques to handle diversity imposed by the requirements and the execution environment, on the other hand we leverage diversity to increase the robustness of software in face of unpredicted situations. More specifically, we address the following V&V challenges.

3.2.4.1. Challenges

One major challenge to build flexible and open yet dependable systems is that current software engineering techniques require architects to foresee all possible situations the system will have to face. However, openness and flexibility also mean unpredictability: unpredictable bugs, attacks, environmental evolutions, etc. Current fault-tolerance [144] and security [115] techniques provide software systems with the capacity of detecting accidental and deliberate faults. However, existing solutions assume that the set of bugs or vulnerabilities in a system does not evolve. This assumption does not hold for open systems, thus it is essential to revisit fault-tolerance and security solutions to account for diverse and unpredictable faults.

Diversity is known to be a major asset for the robustness of large, open, and complex systems (*e.g.*, economical or ecological systems). Following this observation, the software engineering literature provides a rich set of work that choose to implement diversity in software systems in order to improve robustness to attacks or to changes in quality of service. These works range from N-version programming to obfuscation of data structures or control flow, to randomization of instruction sets. An essential remaining challenge is to support the automatic synthesis and evolution of software diversity in open software-intensive systems. There is an opportunity to further enhance these techniques in order to cope with a wider diversity of faults, by multiplying the levels of diversity in the different software layers that are found in software-intensive systems (system, libraries, frameworks, application). This increased diversity must be based on artificial program transformations and code synthesis, which increase the chances of exploring novel solutions, better fitted at one point in time. The biological analogy also indicates that diversity should emerge as a side-effect of evolution, to prevent over-specialization towards one kind of diversity.

3.2.4.2. Scientific objectives

The main objective is to address one of the main limitations of N-version programming for fault-tolerant systems: the manual production and management of software diversity. Through automated injection of artificial diversity we aim at systematically increasing failure diversity and thus increasing the chances of early error detection at run-time. A fundamental assumption for this work is that software-intensive systems can be “good enough” [145], [157].

Proactive program diversification. We aim at establishing novel principles and techniques that favor the emergence of multiple forms of software diversity in software-intensive systems, in conjunction with the software adaptation mechanisms that leverage this diversity. The main expected outcome is a set of meta-design principles that maintain diversity in systems and the experimental demonstration of the effects of software diversity on the adaptive capacities of CASs. Higher levels of diversity in the system provide a pool of software solutions that can eventually be used to adapt to situations unforeseen at design time (bugs, crash, attacks, etc.). Principles of automated software diversification rely on the automated synthesis of variants in a software product line, as well as finer-grained program synthesis combining unsound transformations and genetic programming to explore the space of mutational robustness.

Multi-tier software diversification. We call multi-tier diversification the fact of diversifying several application software components simultaneously. The novelty of our proposal, with respect to the software diversity state of the art, is to diversify the application-level code (for example, diversify the business logics of the application), focusing on the technical layers found in web applications. The diversification of application software

code is expected to provide a diversity of failures and vulnerabilities in web server deployment. Web server deployment usually adopts a form of the Reactor architecture pattern, for scalability purposes: multiple copies of the server software stack, called request handlers, are deployed behind a load balancer. This architecture is very favorable for diversification, since by using the multiplicity of request handlers running in a web server we can simultaneously deploy multiple combinations of diverse software components. Then, if one handler is hacked or crashes the others should still be able to process client requests.

DYOGENE Project-Team

3. Research Program

3.1. Network Calculus

Network calculus [50] is a theory for obtaining deterministic upper bounds in networks that has been developed by R. Cruz [42], [43]. From the modelling point of view, it is an algebra for computing and propagating constraints given in terms of envelopes. A flow is represented by its cumulative function $R(t)$ (that is, the amount of data sent by the flow up to time t). A constraint on a flow is expressed by an arrival curve $\alpha(t)$ that gives an upper bound for the amount of data that can be sent during any interval of length t. Flows cross service elements that offer guarantees on the service. A constraint on a service is a service curve $\beta(t)$ that is used to compute the amount of data that can be served during an interval of length t. It is also possible to define in the same way minimal arrival curves and maximum service curves. Then such constraints envelop the processes and the services. Network calculus enables the following operations:
• computing the exact output cumulative function or at least bounding functions;
• computing output constraints for a flow (like an output arrival curve);
• computing the remaining service curve (that is, the service that of not used by the flows crossing a server);
• composing several servers in tandem;
• giving upper bounds on the worst-case delay and backlog (bounds are tight for a single server or a single flow).
The operations used for this are an adaptation of filtering theory to $(\min, +)$: $(\min, +)$ convolution and deconvolution, sub-additive closure.

We investigate the complexity of computing exact worst-case performance bounds in network calculus and to develop algorithms that present a good trade off between algorithmic efficiency and accuracy of the bounds.

3.2. Perfect Simulation

Simulation approaches can be used to efficiently estimate the stationary behavior of Markov chains by providing independent samples distributed according to their stationary distribution, even when it is impossible to compute this distribution numerically.

The classical Markov Chain Monte Carlo simulation techniques suffer from two main problems:
• The convergence to the stationary distribution can be very slow, and it is in general difficult to estimate;
• Even if one has an effective convergence criterion, the sample obtained after any finite number of iterations is biased.

To overcome these issues, Propp and Wilson [51] have introduced a perfect sampling algorithm (PSA) that has later been extended and applied in various contexts, including statistical physics [46], stochastic geometry [48], theoretical computer science [40], and communications networks [39], [45] (see also the annotated bibliography by Wilson [56]).

Perfect sampling uses coupling arguments to give an unbiased sample from the stationary distribution of an ergodic Markov chain on a finite state space $\mathcal{X}$. Assume the chain is given by an update function Φ and an i.i.d. sequence of innovations $(U_n)_{n\in\mathbb{Z}}$, so that

$$X_{n+1} = \Phi(X_n, U_{n+1}). \tag{89}$$

The algorithm is based on a backward coupling scheme: it computes the trajectories from all $x \in \mathcal{X}$ at some time in the past $t = -T$ until time $t = 0$, using the same innovations. If the final state is the same for all trajectories (i.e. $|\{\Phi(x, U_{-T+1}, ..., U_0) : x \in \mathcal{X}\}| = 1$, where $\Phi(x, U_{-T+1}, ..., U_0) := \Phi(\Phi(x, U_{-T+1}), U_{-T+2}, ..., U_0)$ is defined by induction on T), then we say that the chain has globally coupled and the final state has the stationary distribution of the Markov chain. Otherwise, the simulations are started further in the past.

Any ergodic Markov chain on a finite state space has a representation of type (1) that couples in finite time with probability 1, so Propp and Wilson's PSA gives a "perfect" algorithm in the sense that it provides an *unbiased* sample in *finite time*. Furthermore, the stopping criterion is given by the coupling from the past scheme, and knowing the explicit bounds on the coupling time is not needed for the validity of the algorithm.

However, from the computational side, PSA is efficient only under some monotonicity assumptions that allow reducing the number of trajectories considered in the coupling from the past procedure only to extremal initial conditions. Our goal is to propose new algorithms solving this issue by exploiting semantic and geometric properties of the event space and the state space.

3.3. Stochastic Geometry

Stochastic geometry [54] is a rich branch of applied probability which allows one to quantify random phenomena on the plane or in higher dimension. It is intrinsically related to the theory of point processes. Initially its development was stimulated by applications to biology, astronomy and material sciences. Nowadays it is also widely used in image analysis. It provides a way of estimating and computing "spatial averages". A typical example, with obvious communication implications, is the so called Boolean model, which is defined as the union of discs with random radii (communication ranges) centered at the points of a Poisson point process (user locations) of the Euclidean plane (e.g., a city). A first typical question is that of the prediction of the fraction of the plane which is covered by this union (statistics of coverage). A second one is whether this union has an infinite component or not (connectivity). Further classical models include shot noise processes and random tessellations. Our research consists of analyzing these models with the aim of better understanding wireless communication networks in order to predict and control various network performance metrics. The models require using techniques from stochastic geometry and related fields including point processes, spatial statistics, geometric probability, percolation theory.

3.4. Information Theory

Classical models of stochastic geometry (SG) are not sufficient for analyzing wireless networks as they ignore the specific nature of radio channels.

Consider a wireless communication network made of a collection of nodes which in turn can be transmitters or receivers. At a given time, some subset of this collection of nodes simultaneously transmit, each toward its own receiver. Each transmitter–receiver pair in this snapshot requires its own wireless link. For each such wireless link, the power of the signal received from the link transmitter is jammed by the powers of the signals received from the other transmitters. Even in the simplest model where the power radiated from a point decays in some isotropic way with Euclidean distance, the geometry of the location of nodes plays a key role within this setting since it determines the signal to interference and noise ratio (SINR) at the receiver of each such link and hence the possibility of establishing simultaneously this collection of links at a given bit rate, as shown by information theory (IT). In this definition, the interference seen by some receiver is the sum of the powers of the signals received from all transmitters excepting its own. The SINR field, which is of an essentially geometric nature, hence determines the connectivity and the capacity of the network in a broad sense. The essential point here is that the characteristics and even the feasibilities of the radio links that are simultaneously active are strongly interdependent and determined by the geometry. Our work is centered on the development of an IT-aware stochastic geometry addressing this interdependence.

3.5. The Cavity Method for Network Algorithms

The cavity method combined with geometric networks concepts has recently led to spectacular progresses in digital communications through error-correcting codes. More than fifty years after Shannon's theorems, some coding schemes like turbo codes and low-density parity-check codes (LDPC) now approach the limits predicted by information theory. One of the main ingredients of these schemes is message-passing decoding strategies originally conceived by Gallager, which can be seen as direct applications of the cavity method on a random bipartite graph (with two types of nodes representing information symbols and parity check symbols, see [52]).

Modern coding theory is only one example of application of the cavity method. The concepts and techniques developed for its understanding have applications in theoretical computer science and a rich class of *complex systems*, in the field of networking, economics and social sciences. The cavity method can be used both for the analysis of randomized algorithms and for the study of random ensembles of computational problems representative real-world situations. In order to analyze the performance of algorithms, one generally defines a family of instances and endows it with a probability measure, in the same way as one defines a family of samples in the case of spin glasses or LDPC codes. The discovery that the hardest-to-solve instances, with all existing algorithms, lie close to a *phase transition* boundary has spurred a lot of interest. Theoretical physicists suggest that the reason is a structural one, namely a change in the geometry of the set of solutions related to the *replica symmetry breaking* in the cavity method. Phase transitions, which lie at the core of statistical physics, also play a key role in computer science [53], signal processing [44] and social sciences [47]. Their analysis is a major challenge, that may have a strong impact on the design of related algorithms.

We develop mathematical tools in the theory of discrete probabilities and theoretical computer science in order to contribute to a rigorous formalization of the cavity method, with applications to network algorithms, statistical inference, and at the interface between computer science and economics (EconCS).

3.6. Statistical Learning

Sparse graph structures are useful in a number of information processing tasks where the computational problem can be described as follows: infer the values of a large collection of random variables, given a set of constraints or observations, that induce relations among them. Similar design ideas have been proposed in sensing and signal processing and have applications in coding [41], network measurements, group testing or multi-user detection. While the computational problem is generally hard, sparse graphical structures lead to low-complexity algorithms that are very effective in practice. We develop tools in order to contribute to a precise analysis of these algorithms and of their gap to optimal inference which remains a largely open problem.

A second line of activities concerns the design of protocols and algorithms enabling a transmitter to learn its environment (the statistical properties of the channel quality to the corresponding receiver, as well as their interfering neighbouring transmitters) so as to optimise their transmission strategies and to fairly and efficiently share radio resources. This second objective calls for the development and use of machine learning techniques (e.g. bandit optimisation).

FOCUS Project-Team

3. Research Program

3.1. Models

The objective of Focus is to develop concepts, techniques, and possibly also tools, that may contribute to the analysis and synthesis of CBUS. Fundamental to these activities is *modeling*. Therefore designing, developing and studying computational models appropriate for CBUS is a central activity of the project. The models are used to formalize and verify important computational properties of the systems, as well as to propose new linguistic constructs.

The models we study are in the process calculi (e.g., the π-calculus) and λ-calculus tradition. Such models, with their emphasis on algebra, well address compositionality—a central property in our approach to problems. Accordingly, the techniques we employ are mainly operational techniques based on notions of behavioral equivalence, and techniques based on algebra, mathematical logics, and type theory.

The sections below provide some more details on why process calculi, λ-calculi, and related techniques, should be useful for CBUS.

FUN Project-Team

3. Research Program

3.1. Introduction

The research area of FUN research group is represented in Figure 1 . FUN research group will address every item of Figure 1 starting from the highest level of the figure, *i.e.* in area of homogeneous FUNs to the lowest one. Going down brings more applications and more issues to solve. Results achieved in the upper levels can be re-used in the lower ones. Current networks encountered nowadays are the ones at the higher level, without any interaction between them. In addition, solutions provided for such networks are rarely directly applicable in realistic networks because of the impact of the wireless medium.

FUN research group intends to fill the scientific gap and extend research performed in the area of wireless sensor and actor networks and RFID systems in two directions that are complementary and should be performed in parallel:

- **From theory to experimentation and reciprocally** On one hand, FUN research group intends to investigate new self-organization techniques for these future networks that take into account realistic parameters, emphasizing experimentation and considering mobility.
- **Towards heterogeneous FUNs** On the other hand, FUN research group intends to investigate techniques to allow heterogeneous FUNs to work together in a transparent way for the user. Indeed, new applications integrating several of these components are very much in demand (i.e. smart building) and thus these different technologies need to cooperate.

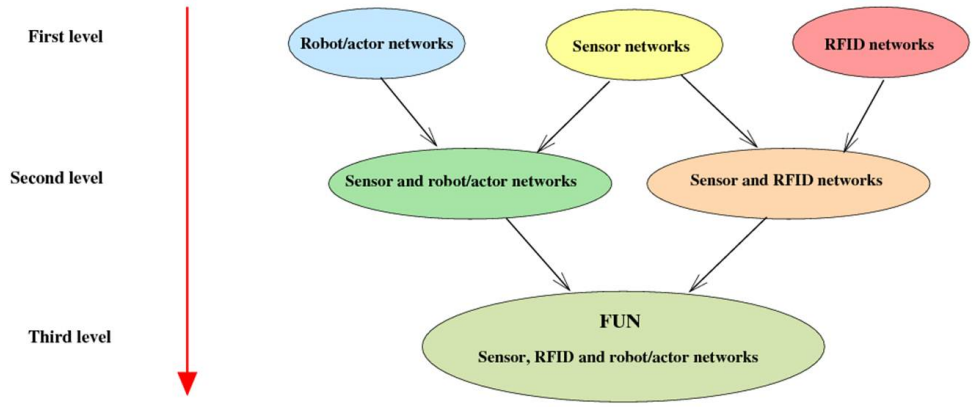


Figure 1. Panorama of FUN.

3.2. From theory to experimentation and reciprocally

Nowadays, even if some powerful and efficient propositions arise in the literature for each of these networks, very few are validated by experimentations. And even when this is the case, no lesson is learnt from it to improve the algorithms. FUN research group needs to study the limits of current assumptions in realistic and mobile environments.

Solutions provided by the FUN research group will mainly be algorithmic. These solutions will first be studied theoretically, principally by using stochastic geometry (like in [47]) or self-stabilization [49] tools in order to derive algorithm behavior in ideal environment. Theory is not an end in itself but only a tool to help in the characterization of the solution in the ideal world. For instance, stochastic geometry will allow quantifying changes in neighborhood or number of hops in a routing path. Self-stabilization will allow measuring stabilization times.

Those same solutions will then be confronted to realistic environments and their 'real' behavior will be analyzed and compared to the expected ones. Comparing theory, simulation and experimentation will allow will allow the influence of a realistic environment be better measured. From this and from the analysis of the information really available for nodes, FUN research group will investigate some means either to counterbalance these effects or to take advantage of them. New solutions provided by the FUN research group will take into consideration the vagaries of a realistic wireless environment and the node mobility. New protocols will take as inputs environmental data (as signal strength or node velocity/position, etc) and node characteristics (the node may have the ability to move in a controlled way) when available. FUN research group will thus adopt a **cross-layered** approach between hardware, physical environment, application requirements, self-organizing and routing techniques. For instance, FUN research group will study how the controlled node mobility can be exploited to enhance the network performance at lowest cost.

Solutions will follow the building process presented by Figure 2 . Propositions will be analyzed not only theoretically and by simulation but also by experimentation to observe the impact of the realistic medium on the behavior of the algorithms. These observations should lead to the derivation of cross-layered models. Experimentation feedbacks will be re-injected in solution design in order to propose algorithms that best fit the environment, and so on till getting satisfactory behavior in both small and large scale environments. All this should be done in such a way that the resulting propositions fit the hardware characteristics (low memory, CPU and energy capacity) and easy to deploy to allow their use by non experts. Since solutions should take into account application requirements as well as hardware characteristics and environment, solutions should be generic enough and then able to self-configure to adapt their environment settings.

In order to achieve this experimental environments, the FUN research group will maintain its strong activity on platform deployment such as SensLAB [52], FIT [25] and Aspire [44]. Next steps will be to experiment not only on testbeds but also on real use cases. These latter will be given through different collaborations.

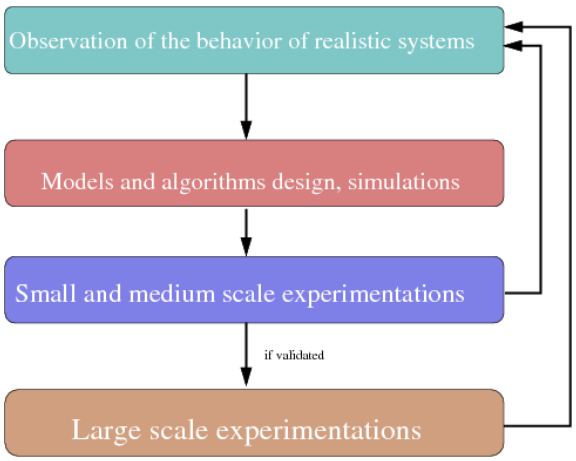


Figure 2. Methodology applied in the FUN research group.

FUN research group will investigate self-organizing techniques for FUNs by providing cross-layered solutions that integrate in their design the adaptability to the realistic environment features. Every solution will be validated with regards to specific application requirements and in realistic environments.

Facing the medium instability. The behavior of wireless propagation is very depending of the surrounding environment (in-door vs outdoor, night vs day, etc) and is very instable. Many experiments in different environment settings should be conducted. Experiment platforms such as SensLAB, FIT, our wifiBot as robots and actuators and our RFID devices will be used offering ways to experiment easily and quickly in different environments but might not be sufficient to experiment every environment.
Adaptability and flexibility. Since from one application to another one, requirements and environments are different, solutions provided by FUN research group should be **generic** enough and **self-adapt** to their environment. Algorithm design and validation should also take into account the targeted applications brought for instance by our industrial partners like Etineo. All solution designs should keep in mind the devices constrained capacities. Solutions should consume low resources in terms of memory, processor and energy to provide better performances and scale. All should be self-adaptive.

FUN research group will try to take advantage of some observed features that could first be seen as drawbacks. For instance, the broadcast nature of wireless networks is first an inconvenient since the use of a link between two nodes inhibits every other communication in the same transmission area. But algorithms should exploit that feature to derive new behaviors and a node blocked by another transmission should overhear it to get more information and maybe to limit the overall information to store in the network or overhead communication.

3.3. Towards unified heterogeneous FUNs

The second main direction to be followed by the FUN research group is to merge networks from the upper layer in Fig. 1 into networks from the lowest level. Indeed, nowadays, these networks are still considered as separated issues. But considering mixed networks bring new opportunities. Indeed, robots can deploy, replace, compensate sensor nodes. They also can collect periodically their data, which avoids some long and multi-hop communications between sensor nodes and thus preserving their resources. Robots can also perform many additional tasks to enhance network performance like positioning themselves on strategic points to ensure area coverage or reduce routing path lengths. Similarly, coupling sensors and RFID tags also bring new opportunities that are more and more in-demand from the industrial side. Indeed, an RFID reader may be a sensor in a wireless sensor network and data hold by RFID tags and collected by readers might need to be reported to a sink. This will allow new applications and possibilities such as the localization of a tagged object in an environment be covered by sensors.

When at last all components are gathered, this leads us to a new era in which every object is autonomous. Let's consider for instance a smart home equipped with sensors and RFID reader. An event triggered by a sensor (*i.e.* an increase of the temperature) or a RFID reader (*i.e.* detection of a tag hold by a person) will trigger actions from actuators (*i.e.* lowering of stores, door opening). Possibilities are huge. But with all these new opportunities come new technological issues with other constraints. Every entity is considered as an object possibly mobile which should be dynamically identified and controlled. To support this dynamics, protocols should be localized and distributed. Model derived from experiment observations should be unified to fit all these classes of devices.

FUN research group will investigate new protocols and communication paradigms that allow the technologies to be transparently merged. Objects and events might interconnect while respecting on-going standards and building an autonomic and smart network while being compliant with hardware resources and environment.

Technologies such as wireless sensors, wireless robots/actuators and RFID tags/ readers, although presenting many common points are still part of different disciplines that have evolved in parallel ways. Every branch is at different maturity levels and has developed its own standards. Nevertheless, making all these devices part of a single unified network leverages technological issues (partly addressed in the former objective) but also regarding to on-going standards and data formatting. FUN research group will have to study current standards

of every area in order to propose compliant solutions. Such works have been initiated in the POPS research group in the framework of the FP7 ASPIRE project. Members of FUN research group intend to continue and enlarge these works.

Todays' EPCGlobal compliant RFID readers must comply to some rules and be configurable through an ALE (Application Level Event) [42]. While a fixed and connected RFID reader is easily configurable, configuring remotely a mobile RFID reader might be very difficult since it implies to first locate it and then send configuration data through a wireless dynamic network. FUN research group will investigate some tools that make the configuration easy and transparent for the user. This remote configuration of mobile readers through the network should consider application requirements and network and reader characteristics to choose the best trade-off relative to the software part embedded in the reader. The biggest part embedded, the lowest bandwidth overhead (data can be filtered and aggregated in the reader) and the greater mobility (readers are still fully operational even when disconnected) but the more difficult to set up and the more powerful readers. All these aspects will be studied within the FUN research group.

GANG Project-Team

3. Research Program

3.1. Graph and Combinatorial Algorithms

We focus on two approaches for designing algorithms for large graphs: decomposing the graph and and relying on simple graph traversals.

3.1.1. Graph Decompositions

We study new decompositions schemes such as 2-join, skew partitions and others partition problems. These graph decompositions appeared in the structural graph theory and are the basis of some well-known theorems such as the Perfect Graph Theorem. For these decompositions there is a lack of efficient algorithms. We aim at designing algorithms working in $O(nm)$ since we think that this could be a lower bound for these decompositions.

3.1.2. Graph Search

We more deeply study multi-sweep graph searches. In this domain a graph search only yields a total ordering of the vertices which can be used by the subsequent graph searches. This technique can be used on huge graphs and do not need extra memory. We already have obtained preliminary results in this direction and many well-known graph algorithms can be put in this framework. The idea behind this approach is that each sweep discovers some structure of the graph. At the end of the process either we have found the underlying structure (for example an interval representation for an interval graph) or an approximation of it (for example in hard discrete optimization problems). We envision applications to exact computations of centers in huge graphs, to underlied combinatorial optimization problems, but also to networks arising in Biology.

3.1.3. Graph Exploration

In the course of graph exploration, a mobile agent is expected to regularly visit all the nodes of an unknown network, trying to discover all its nodes as quickly as possible. Our research focuses on the design and analysis of agent-based algorithms for exploration-type problems, which operate efficiently in a dynamic network environment, and satisfy imposed constraints on local computational resources, performance, and resilience. Our recent contributions in this area concern the design of fast deterministic algorithms for teams of agents operating in parallel in a graph, with limited or no persistent state information available at nodes. We plan further studies to better understand the impact of memory constraints and of the availability of true randomness on efficiency of the graph exploration process.

3.2. Distributed Computing

The distributed community can be viewed as the union of two sub-communities. This is true even in our team. Even though they are not completely disjoint, they are disjoint enough not to leverage each other's results. At a high level, one is mostly interested in timing issues (clock drifts, link delays, crashes, etc.) while the other one is mostly interested in spatial issues (network structure, memory requirements, etc.). Indeed, one sub-community is mostly focusing on the combined impact of asynchronism and faults on distributed computation, while the other addresses the impact of network structural properties on distributed computation. Both communities address various forms of computational complexities, through the analysis of different concepts. This includes, e.g., failure detectors and wait-free hierarchy for the former community, and compact labeling schemes and computing with advice for the latter community. We have the ambitious project to achieve the reconciliation between the two communities by focusing on the same class of problems, the yes/no-problems, and establishing the scientific foundations for building up a consistent theory of computability and complexity for distributed computing. The main question addressed is therefore: is the absence of globally coherent computational complexity theories covering more than fragments of distributed computing, inherent

to the field? One issue is obviously the types of problems located at the core of distributed computing. Tasks like consensus, leader election, and broadcasting are of very different nature. They are not *yes-no* problems, neither are they minimization problems. Coloring and Minimal Spanning Tree are optimization problems but we are often more interested in constructing an optimal solution than in verifying the correctness of a given solution. Still, it makes full sense to analyze the *yes-no* problems corresponding to checking the validity of the output of tasks. Another issue is the power of individual computation. The FLP impossibility result as well as Linial's lower bound hold independently from the individual computational power of the involved computing entities. For instance, the individual power of solving NP-hard problems in constant time would not help overcoming these limits which are inherent to the fact that computation is distributed. A third issue is the abundance of models for distributed computing frameworks, from shared memory to message passing, spanning all kinds of specific network structures (complete graphs, unit-disk graphs, etc.) and or timing constraints (from complete synchronism to full asynchronism). There are however models, typically the wait-free model and the LOCAL model, which, though they do not claim to reflect accurately real distributed computing systems, enable focusing on some core issues. Our research program is ongoing to carry many important notions of Distributed Computing into a *standard* computational complexity.

3.3. Network Algorithms and Analysis

Based on our scientific foundation on both graph algorithms and distributed algorithms, we plan to analyze the behavior of various networks such as future Internet, social networks, overlay networks resulting from distributed applications or online social networks.

3.3.1. Information Dissemination

One of the key aspects of networks resides in the dissemination of information among the nodes. We aim at analyzing various procedures of information propagation from dedicated algorithms to simple distributed schemes such as flooding. We also consider various models, where noise can alter information as it propagates or where memory of nodes is limited for example.

3.3.2. Routing Paradigms

We try to explore new routing paradigms such as greedy routing in social networks for example. We are also interested in content centric networking where routing is based on content name rather than content address. One of our target is multiple path routing: how to design forwarding tables providing multiple disjoint paths to a destination?

3.3.3. Beyond Peer-to-Peer

Based on our past experience of peer-to-peer application design, we would like to broaden the spectrum of distributed applications where new efficient algorithms and analysis can be performed. We especially target online social networks if we see them as collaborative tools for exchanging information. A basic question resides in making the right connections for gathering filtered and accurate information with sufficient coverage.

3.3.4. SAT and Forwarding Information Verification

As forwarding tables of networks grow and are sometimes manually modified, the problem of verifying forwarding information becomes critical and has recently gained in interest. Some problems that arise in network verification such as loop detection for example, may be naturally encoded as Boolean Satisfiability problems. Beside the theoretical interest of this encoding in complexity proofs, it has also a practical value for solving these problems by taking advantage of the many efficient Satisfiability testing solvers. Indeed, SAT solvers have proved to be very efficient in solving problems coming from various areas (Circuit Verification, Dependency and Conflicts in Software distributions...) and encoded in Conjunctive Normal Form. To test an approach using SAT solvers in network verification, one need to collect data sets from real network and to develop good models for generating realistic networks. The technique of encoding and the solvers themselves need to be adapted to this kind of problems. All this represent a rich experimental field of future research.

3.3.5. Network Analysis

Finally, we are interested in analyzing the structural properties of practical networks. This can include diameter computation or ranking of nodes. As we mostly consider large networks, we are often interested in efficient heuristics. Ideally, we target heuristics that give exact answer although fast computation time is not guaranteed for all networks. We already have designed such heuristics for diameter computation; understanding the structural properties that enable fast computation time in practice is still an open question.

HIEPACS Project-Team

3. Research Program

3.1. Introduction

The methodological component of HIEPACS concerns the expertise for the design as well as the efficient and scalable implementation of highly parallel numerical algorithms to perform frontier simulations. In order to address these computational challenges a hierarchical organization of the research is considered. In this bottom-up approach, we first consider in Section 3.2 generic topics concerning high performance computational science. The activities described in this section are transversal to the overall project and their outcome will support all the other research activities at various levels in order to ensure the parallel scalability of the algorithms. The aim of this activity is not to study general purpose solution but rather to address these problems in close relation with specialists of the field in order to adapt and tune advanced approaches in our algorithmic designs. The next activity, described in Section 3.3 , is related to the study of parallel linear algebra techniques that currently appear as promising approaches to tackle huge problems on extreme scale platforms. We highlight the linear problems (linear systems or eigenproblems) because they are in many large scale applications the main computational intensive numerical kernels and often the main performance bottleneck. These parallel numerical techniques, which are involved in the IPL C2S@EXA, will be the basis of both academic and industrial collaborations, some are described in Section 4.1 , but will also be closely related to some functionalities developed in the parallel fast multipole activity described in Section 3.4 . Finally, as the accuracy of the physical models increases, there is a real need to go for parallel efficient algorithm implementation for multiphysics and multiscale modeling in particular in the context of code coupling. The challenges associated with this activity will be addressed in the framework of the activity described in Section 3.5 .

Currently, we have one major application (see Section 4.1) that is in material physics. We will contribute to all steps of the design of the parallel simulation tool. More precisely, our applied mathematics skill will contribute to the modelling, our advanced numerical schemes will help in the design and efficient software implementation for very large parallel multi-scale simulations. We also participate to a few co-design actions in close collaboration with some applicative groups. The objective of this activity is to instanciate our expertise in fields where they are critical for designing scalable simulation tools. We refer to Section 4.2 for a detailed description of these activities.

3.2. High-performance computing on next generation architectures

Participants: Emmanuel Agullo, Olivier Coulaud, Luc Giraud, Mathieu Faverge, Abdou Guermouche, Matías Hastaran, Andra Hugo, Xavier Lacoste, Guillaume Latu, Stojce Nakov, Florent Pruvost, Pierre Ramet, Jean Roman, Mawussi Zounon.

The research directions proposed in HIEPACS are strongly influenced by both the applications we are studying and the architectures that we target (i.e., massively parallel many-core architectures, ...). Our main goal is to study the methodology needed to efficiently exploit the new generation of high-performance computers with all the constraints that it induces. To achieve this high-performance with complex applications we have to study both algorithmic problems and the impact of the architectures on the algorithm design.

From the application point of view, the project will be interested in multiresolution, multiscale and hierarchical approaches which lead to multi-level parallelism schemes. This hierarchical parallelism approach is necessary to achieve good performance and high-scalability on modern massively parallel platforms. In this context, more specific algorithmic problems are very important to obtain high performance. Indeed, the kind of applications we are interested in are often based on data redistribution for example (e.g. code coupling applications). This well-known issue becomes very challenging with the increase of both the number of computational nodes and the amount of data. Thus, we have both to study new algorithms and to adapt the

existing ones. In addition, some issues like task scheduling have to be restudied in this new context. It is important to note that the work done in this area will be applied for example in the context of code coupling (see Section 3.5).

Considering the complexity of modern architectures like massively parallel architectures or new generation heterogeneous multicore architectures, task scheduling becomes a challenging problem which is central to obtain a high efficiency. Of course, this work requires the use/design of scheduling algorithms and models specifically to tackle our target problems. This has to be done in collaboration with our colleagues from the scheduling community like for example O. Beaumont (Inria REALOPT Project-Team). It is important to note that this topic is strongly linked to the underlying programming model. Indeed, considering multicore architectures, it has appeared, in the last five years, that the best programming model is an approach mixing multi-threading within computational nodes and message passing between them. In the last five years, a lot of work has been developed in the high-performance computing community to understand what is critic to efficiently exploit massively multicore platforms that will appear in the near future. It appeared that the key for the performance is firstly the grain of computations. Indeed, in such platforms the grain of the parallelism must be small so that we can feed all the processors with a sufficient amount of work. It is thus very crucial for us to design new high performance tools for scientific computing in this new context. This will be developed in the context of our solvers, for example, to adapt to this new parallel scheme. Secondly, the larger the number of cores inside a node, the more complex the memory hierarchy. This remark impacts the behaviour of the algorithms within the node. Indeed, on this kind of platforms, NUMA effects will be more and more problematic. Thus, it is very important to study and design data-aware algorithms which take into account the affinity between computational threads and the data they access. This is particularly important in the context of our high-performance tools. Note that this work has to be based on an intelligent cooperative underlying run-time (like the tools developed by the Inria RUNTIME Project-Team) which allows a fine management of data distribution within a node.

Another very important issue concerns high-performance computing using “heterogeneous” resources within a computational node. Indeed, with the emergence of the GPU and the use of more specific co-processors, it is important for our algorithms to efficiently exploit these new kind of architectures. To adapt our algorithms and tools to these accelerators, we need to identify what can be done on the GPU for example and what cannot. Note that recent results in the field have shown the interest of using both regular cores and GPU to perform computations. Note also that in opposition to the case of the parallelism granularity needed by regular multicore architectures, GPU requires coarser grain parallelism. Thus, making both GPU and regular cores work all together will lead to two types of tasks in terms of granularity. This represents a challenging problem especially in terms of scheduling. From this perspective, we investigate new approaches for composing parallel applications within a runtime system for heterogeneous platforms.

The SOLHAR project aims at studying and designing algorithms and parallel programming models for implementing direct methods for the solution of sparse linear systems on emerging computers equipped with accelerators. Several attempts have been made to accomplish the porting of these methods on such architectures; the proposed approaches are mostly based on a simple offloading of some computational tasks (the coarsest grained ones) to the accelerators and rely on fine hand-tuning of the code and accurate performance modeling to achieve efficiency. SOLHAR proposes an innovative approach which relies on the efficiency and portability of runtime systems, such as the StarPU tool developed in the RUNTIME team. Although the SOLHAR project will focus on heterogeneous computers equipped with GPUs due to their wide availability and affordable cost, the research accomplished on algorithms, methods and programming models will be readily applicable to other accelerator devices. Our final goal would be to have high performance solvers and tools which can efficiently run on all these types of complex architectures by exploiting all the resources of the platform (even if they are heterogeneous).

In order to achieve an advanced knowledge concerning the design of efficient computational kernels to be used on our high performance algorithms and codes, we will develop research activities first on regular frameworks before extending them to more irregular and complex situations. In particular, we will work first on optimized dense linear algebra kernels and we will use them in our more complicated direct and hybrid

solvers for sparse linear algebra and in our fast multipole algorithms for interaction computations. In this context, we will participate to the development of those kernels in collaboration with groups specialized in dense linear algebra. In particular, we intend develop a strong collaboration with the group of Jack Dongarra at the University of Tennessee and collaborating research groups. The objectives will be to develop dense linear algebra algorithms and libraries for multicore architectures in the context the PLASMA project and for GPU and hybrid multicore/GPU architectures in the context of the MAGMA project. The framework that hosts all these research activities is the associate team MORSE.

A more prospective objective is to study the resiliency in the context of large-scale scientific applications for massively parallel architectures. Indeed, with the increase of the number of computational cores per node, the probability of a hardware crash on a core or of a memory corruption is dramatically increased. This represents a crucial problem that needs to be addressed. However, we will only study it at the algorithmic/application level even if it needed lower-level mechanisms (at OS level or even hardware level). Of course, this work can be performed at lower levels (at operating system) level for example but we do believe that handling faults at the application level provides more knowledge about what has to be done (at application level we know what is critical and what is not). The approach that we will follow will be based on the use of a combination of fault-tolerant implementations of the run-time environments we use (like for example FT-MPI) and an adaptation of our algorithms to try to manage this kind of faults. This topic represents a very long range objective which needs to be addressed to guaranty the robustness of our solvers and applications. In that respect, we are involved in a ANR-Blanc project entitles RESCUE jointly with two other Inria EPI, namely ROMA and GRAND-LARGE and the G8 ESC international initiative as well as in the EXA2CT FP7 project. The main objective of the RESCUE project is to develop new algorithmic techniques and software tools to solve the exascale resilience problem. Solving this problem implies a departure from current approaches, and calls for yet-to-be- discovered algorithms, protocols and software tools.

Finally, it is important to note that the main goal of HIEPACS is to design tools and algorithms that will be used within complex simulation frameworks on next-generation parallel machines. Thus, we intend with our partners to use the proposed approach in complex scientific codes and to validate them within very large scale simulations as well as designing parallel solution in co-design collaborations.

3.3. High performance solvers for large linear algebra problems

Participants: Emmanuel Agullo, Astrid Casadei, Olivier Coulaud, Mathieu Faverge, Romain Garnier, Luc Giraud, Abdou Guermouche, Andra Hugo, Xavier Lacoste, Pablo Salas Medina, Stojce Nakov, Julien Pedron, Florent Pruvost, Pierre Ramet, Jean Roman.

Starting with the developments of basic linear algebra kernels tuned for various classes of computers, a significant knowledge on the basic concepts for implementations on high-performance scientific computers has been accumulated. Further knowledge has been acquired through the design of more sophisticated linear algebra algorithms fully exploiting those basic intensive computational kernels. In that context, we still look at the development of new computing platforms and their associated programming tools. This enables us to identify the possible bottlenecks of new computer architectures (memory path, various level of caches, inter processor or node network) and to propose ways to overcome them in algorithmic design. With the goal of designing efficient scalable linear algebra solvers for large scale applications, various tracks will be followed in order to investigate different complementary approaches. Sparse direct solvers have been for years the methods of choice for solving linear systems of equations, it is nowadays admitted that classical approaches are not scalable neither from a computational complexity nor from a memory view point for large problems such as those arising from the discretization of large 3D PDE problems. We will continue to work on sparse direct solvers on the one hand to make sure they fully benefit from most advanced computing platforms and on the other hand to attempt to reduce their memory and computational costs for some classes of problems where data sparse ideas can be considered. Furthermore, sparse direct solvers are a key building boxes for the design of some of our parallel algorithms such as the hybrid solvers described in the sequel of this section. Our activities in that context will mainly address preconditioned Krylov subspace methods; both components, preconditioner and Krylov solvers, will be investigated. In this framework, and possibly in relation with the

research activity on fast multipole, we intend to study how emerging $\mathcal{H}$-matrix arithmetic can benefit to our solver research efforts.

3.3.1. Parallel sparse direct solver

Solving large sparse systems $Ax = b$ of linear equations is a crucial and time-consuming step, arising in many scientific and engineering applications. Consequently, many parallel techniques for sparse matrix factorization have been studied and implemented.

Sparse direct solvers are mandatory when the linear system is very ill-conditioned; such a situation is often encountered in structural mechanics codes, for example. Therefore, to obtain an industrial software tool that must be robust and versatile, high-performance sparse direct solvers are mandatory, and parallelism is then necessary for reasons of memory capability and acceptable solution time. Moreover, in order to solve efficiently 3D problems with more than 50 million unknowns, which is now a reachable challenge with new multicore supercomputers, we must achieve good scalability in time and control memory overhead. Solving a sparse linear system by a direct method is generally a highly irregular problem that induces some challenging algorithmic problems and requires a sophisticated implementation scheme in order to fully exploit the capabilities of modern supercomputers.

New supercomputers incorporate many microprocessors which are composed of one or many computational cores. These new architectures induce strongly hierarchical topologies. These are called NUMA architectures. In the context of distributed NUMA architectures, in collaboration with the Inria RUNTIME team, we study optimization strategies to improve the scheduling of communications, threads and I/O. We have developed dynamic scheduling designed for NUMA architectures in the PaStiX solver. The data structures of the solver, as well as the patterns of communication have been modified to meet the needs of these architectures and dynamic scheduling. We are also interested in the dynamic adaptation of the computation grain to use efficiently multi-core architectures and shared memory. Experiments on several numerical test cases have been performed to prove the efficiency of the approach on different architectures.

In collaboration with the ICL team from the University of Tennessee, and the RUNTIME team from Inria, we are evaluating the way to replace the embedded scheduling driver of the PaStiX solver by one of the generic frameworks, PaRSEC or StarPU, to execute the task graph corresponding to a sparse factorization. The aim is to design algorithms and parallel programming models for implementing direct methods for the solution of sparse linear systems on emerging computer equipped with GPU accelerators. More generally, this work will be performed in the context of the associate team MORSE and the ANR SOLHAR project which aims at designing high performance sparse direct solvers for modern heterogeneous systems. This ANR project involves several groups working either on the sparse linear solver aspects (HIEPACS and ROMA from Inria and APO from IRIT), on runtime systems (RUNTIME from Inria) or scheduling algorithms (REALOPT and ROMA from Inria). The results of these efforts will be validated in the applications provided by the industrial project members, namely CEA-CESTA and Airbus Group Innovations.

On the numerical side, we are studying how the data sparsness that might exist in some dense blocks appearing during the factorization can be exploited using different compression techniques based on $\mathcal{H}$-matrix (and variants) arithmetics. This research activity will be conducted in the framework of the FASTLA associate team and will naturally irrigate the hybrid solvers described below as well as closely interact with the sparse direct solver actions as well as the other research efforts where similar data sparsness might be exploited.

3.3.2. Hybrid direct/iterative solvers based on algebraic domain decomposition techniques

One route to the parallel scalable solution of large sparse linear systems in parallel scientific computing is the use of hybrid methods that hierarchically combine direct and iterative methods. These techniques inherit the advantages of each approach, namely the limited amount of memory and natural parallelization for the iterative component and the numerical robustness of the direct part. The general underlying ideas are not new since they have been intensively used to design domain decomposition techniques; those approaches cover a fairly large range of computing techniques for the numerical solution of partial differential equations (PDEs) in time and space. Generally speaking, it refers to the splitting of the computational domain into sub-domains with or

without overlap. The splitting strategy is generally governed by various constraints/objectives but the main one is to express parallelism. The numerical properties of the PDEs to be solved are usually intensively exploited at the continuous or discrete levels to design the numerical algorithms so that the resulting specialized technique will only work for the class of linear systems associated with the targeted PDE.

In that context, we intend to continue our effort on the design of algebraic non-overlapping domain decomposition techniques that rely on the solution of a Schur complement system defined on the interface introduced by the partitioning of the adjacency graph of the sparse matrix associated with the linear system. Although it is better conditioned than the original system the Schur complement needs to be precondition to be amenable to a solution using a Krylov subspace method. Different hierarchical preconditioners will be considered, possibly multilevel, to improve the numerical behaviour of the current approaches implemented in our software libraries HIPS and MaPHyS. This activity will be developed in the context of the ANR DEDALES project. In addition to this numerical studies, advanced parallel implementation will be developed that will involve close collaborations between the hybrid and sparse direct activities.

3.3.3. Linear Krylov solvers

Preconditioning is the main focus of the two activities described above. They aim at speeding up the convergence of a Krylov subspace method that is the complementary component involved in the solvers of interest for us. In that framework, we believe that various aspects deserve to be investigated; we will consider the following ones:

- preconditioned block Krylov solvers for multiple right-hand sides. In many large scientific and industrial applications, one has to solve a sequence of linear systems with several right-hand sides given simultaneously or in sequence (radar cross section calculation in electromagnetism, various source locations in seismic, parametric studies in general, ...). For “simultaneous" right-hand sides, the solvers of choice have been for years based on matrix factorizations as the factorization is performed once and simple and cheap block forward/backward substitutions are then performed. In order to effectively propose alternative to such solvers, we need to have efficient preconditioned Krylov subspace solvers. In that framework, block Krylov approaches, where the Krylov spaces associated with each right-hand side are shared to enlarge the search space will be considered. They are not only attractive because of this numerical feature (larger search space), but also from an implementation point of view. Their block-structures exhibit nice features with respect to data locality and re-usability that comply with the memory constraint of multicore architectures. Following the initial work by J. Yan Fei during his post-doc in HIEPACS, we will continue the numerical study of the block GMRES variant that combines inexact break-down detection and deflation at restart. In addition a special attention will be paid to situations where a massive number of right-hand sides are given where variants exploiting the possible sparsness (i.e., compression using $\mathcal{H}$-matrix arithmetic) of these right-hand sides will be explored to design efficient numerical algorithms. Beyond new numerical investigations, a software implementation to be included in our linear solver libray will be developed in the context of the DGA HIBOX project.

  For right-hand sides available one after each other, various strategies that exploit the information available in the sequence of Krylov spaces (e.g. spectral information) will be considered that include for instance technique to perform incremental update of the preconditioner or to build augmented Krylov subspaces.
- Extension or modification of Krylov subspace algorithms for multicore architectures: finally to match as much as possible to the computer architecture evolution and get as much as possible performance out of the computer, a particular attention will be paid to adapt, extend or develop numerical schemes that comply with the efficiency constraints associated with the available computers. Nowadays, multicore architectures seem to become widely used, where memory latency and bandwidth are the main bottlenecks; investigations on communication avoiding techniques will be undertaken in the framework of preconditioned Krylov subspace solvers as a general guideline for all the items mentioned above. This research activity will benefit from the FP7 EXA2CT project led by HIEPACS on behalf of the IPL C2S@EXA that involves two other Inria projects namely

ALPINES and SAGE.

3.3.4. Eigensolvers

Many eigensolvers also rely on Krylov subspace techniques. Naturally some links exist between the Krylov subspace linear solvers and the Krylov subspace eigensolvers. We plan to study the computation of eigenvalue problems with respect to the following two different axes:

- Exploiting the link between Krylov subspace methods for linear system solution and eigensolvers, we intend to develop advanced iterative linear methods based on Krylov subspace methods that use some spectral information to build part of a subspace to be recycled, either though space augmentation or through preconditioner update. This spectral information may correspond to a certain part of the spectrum of the original large matrix or to some approximations of the eigenvalues obtained by solving a reduced eigenproblem. This technique will also be investigated in the framework of block Krylov subspace methods.
- In the context of the calculation of the ground state of an atomistic system, eigenvalue computation is a critical step; more accurate and more efficient parallel and scalable eigensolvers are required.

3.4. High performance Fast Multipole Method for N-body problems

Participants: Emmanuel Agullo, Bérenger Bramas, Arnaud Etcheverry, Olivier Coulaud, Matthias Messner, Cyrille Piacibello, Guillaume Sylvand.

In most scientific computing applications considered nowadays as computational challenges (like biological and material systems, astrophysics or electromagnetism), the introduction of hierarchical methods based on an octree structure has dramatically reduced the amount of computation needed to simulate those systems for a given accuracy. For instance, in the N-body problem arising from these application fields, we must compute all pairwise interactions among N objects (particles, lines, ...) at every timestep. Among these methods, the Fast Multipole Method (FMM) developed for gravitational potentials in astrophysics and for electrostatic (coulombic) potentials in molecular simulations solves this N-body problem for any given precision with $O(N)$ runtime complexity against $O(N^2)$ for the direct computation.

The potential field is decomposed in a near field part, directly computed, and a far field part approximated thanks to multipole and local expansions. We introduced a matrix formulation of the FMM that exploits the cache hierarchy on a processor through the Basic Linear Algebra Subprograms (BLAS). Moreover, we developed a parallel adaptive version of the FMM algorithm for heterogeneous particle distributions, which is very efficient on parallel clusters of SMP nodes. Finally on such computers, we developed the first hybrid MPI-thread algorithm, which enables to reach better parallel efficiency and better memory scalability. We plan to work on the following points in HIEPACS.

3.4.1. Improvement of calculation efficiency

Nowadays, the high performance computing community is examining alternative architectures that address the limitations of modern cache-based designs. GPU (Graphics Processing Units) and the Cell processor have thus already been used in astrophysics and in molecular dynamics. The Fast Mutipole Method has also been implemented on GPU. We intend to examine the potential of using these forthcoming processors as a building block for high-end parallel computing in N-body calculations. More precisely, we want to take advantage of our specific underlying BLAS routines to obtain an efficient and easily portable FMM for these new architectures. Algorithmic issues such as dynamic load balancing among heterogeneous cores will also have to be solved in order to gather all the available computation power. This research action will be conduced on close connection with the activity described in Section 3.2 .

3.4.2. Non uniform distributions

In many applications arising from material physics or astrophysics, the distribution of the data is highly non uniform and the data can grow between two time steps. As mentioned previously, we have proposed a hybrid MPI-thread algorithm to exploit the data locality within each node. We plan to further improve the load balancing for highly non uniform particle distributions with small computation grain thanks to dynamic load balancing at the thread level and thanks to a load balancing correction over several simulation time steps at the process level.

3.4.3. Fast multipole method for dislocation operators

The engine that we develop will be extended to new potentials arising from material physics such as those used in dislocation simulations. The interaction between dislocations is long ranged ($O(1/r)$) and anisotropic, leading to severe computational challenges for large-scale simulations. Several approaches based on the FMM or based on spatial decomposition in boxes are proposed to speed-up the computation. In dislocation codes, the calculation of the interaction forces between dislocations is still the most CPU time consuming. This computation has to be improved to obtain faster and more accurate simulations. Moreover, in such simulations, the number of dislocations grows while the phenomenon occurs and these dislocations are not uniformly distributed in the domain. This means that strategies to dynamically balance the computational load are crucial to achieve high performance.

3.4.4. Fast multipole method for boundary element methods

The boundary element method (BEM) is a well known solution of boundary value problems appearing in various fields of physics. With this approach, we only have to solve an integral equation on the boundary. This implies an interaction that decreases in space, but results in the solution of a dense linear system with $O(N^3)$ complexity. The FMM calculation that performs the matrix-vector product enables the use of Krylov subspace methods. Based on the parallel data distribution of the underlying octree implemented to perform the FMM, parallel preconditioners can be designed that exploit the local interaction matrices computed at the finest level of the octree. This research action will be conduced on close connection with the activity described in Section 3.3 . Following our earlier experience, we plan to first consider approximate inverse preconditionners that can efficiently exploit these data structures.

3.5. Efficient algorithmic for load balancing and code coupling in complex simulations

Participants: Astrid Casadei, Olivier Coulaud, Aurélien Esnard, Maria Predari, Pierre Ramet, Jean Roman.

Many important physical phenomena in material physics and climatology are inherently complex applications. They often use multi-physics or multi-scale approaches, that couple different models and codes. The key idea is to reuse available legacy codes through a coupling framework instead of merging them into a standalone application. There is typically one model per different scale or physics; and each model is implemented by a parallel code. For instance, to model a crack propagation, one uses a molecular dynamic code to represent the atomistic scale and an elasticity code using a finite element method to represent the continuum scale. Indeed, fully microscopic simulations of most domains of interest are not computationally feasible. Combining such different scales or physics is still a challenge to reach high performance and scalability. If the model aspects are often well studied, there are several open algorithmic problems, that we plan to investigate in the HIEPACS project-team.

3.5.1. Efficient schemes for multiscale simulations

As mentioned previously, many important physical phenomena, such as material deformation and failure (see Section 4.1), are inherently multiscale processes that cannot always be modeled via continuum model. Fully microscopic simulations of most domains of interest are not computationally feasible. Therefore, researchers must look at multiscale methods that couple micro models and macro models. Combining different scales such as quantum-atomistic or atomistic, mesoscale and continuum, are still a challenge to obtain efficient and

accurate schemes that efficiently and effectively exchange information between the different scales. We are currently involved in two national research projects, that focus on multiscale schemes. More precisely, the models that we start to study are the quantum to atomic coupling (QM/MM coupling) in the ANR NOSSI and the atomic to dislocation coupling in the ANR OPTIDIS.

3.5.2. Dynamic load balancing for massively parallel coupled codes

In this context of code coupling, one crucial issue is undoubtedly the load balancing of the whole coupled simulation that remains an open question. The goal here is to find the best data distribution for the whole coupled simulation and not only for each standalone code, as it is most usually done. Indeed, the naive balancing of each code on its own can lead to an important imbalance and to a communication bottleneck during the coupling phase, that can drastically decrease the overall performance. Therefore, one argues that it is required to model the coupling itself in order to ensure a good scalability, especially when running on massively parallel architectures (tens of thousands of processors/cores). In other words, one must develop new algorithms and software implementation to perform a *coupling-aware* partitioning of the whole application.

Another related problem is the problem of resource allocation. This is particularly important for the global coupling efficiency and scalabilty, because each code involved in the coupling can be more or less computationally intensive, and there is a good trade-off to find between resources assigned to each code to avoid that one of them waits for the other(s). And what happens if the load of one code dynamically changes relatively to the other? In such a case, it could be convenient to dynamically adapt the number of resources used at runtime.

For instance, the conjugate heat transfer simulation in complex geometries (as developed by the CFD team of CERFACS) requires to couple a fluid/convection solver (AVBP) with a solid/conduction solver (AVTP). The AVBP code is much more CPU consuming than the AVTP code. As a consequence, there is an important computational imbalance between the two solvers. The use of new algorithms to correctly load balance coupled simulations with enhanced graph partitioning techniques appears as a promising way to reach better performances of coupled application on massively parallel computers.

3.5.3. Graph partitioning for hybrid solvers

Graph handling and partitioning play a central role in the activity described here but also in other numerical techniques detailed in Section 3.3 .

The Nested Dissection is now a well-known heuristic for sparse matrix ordering to both reduce the fill-in during numerical factorization and to maximize the number of independent computation tasks. By using the block data structure induced by the partition of separators of the original graph, very efficient parallel block solvers have been designed and implemented according to supernodal or multifrontal approaches. Considering hybrid methods mixing both direct and iterative solvers such as `HIPS` or `MaPHyS`, obtaining a domain decomposition leading to a good balancing of both the size of domain interiors and the size of interfaces is a key point for load balancing and efficiency in a parallel context. We intend to revisit some well-known graph partitioning techniques in the light of the hybrid solvers and design new algorithms to be tested in the `Scotch` package.

HIPERCOM2 Team

3. Research Program

3.1. Methodology of telecommunication algorithm evaluation

We develop our performance evaluation tools towards deterministic performance and probabilistic performance. Our tools range from mathematical analysis to simulation and real life experiment of telecommunication algorithms.

One cannot design good algorithms without good evaluation models. Hipercom project team has an historically strong experience in performance evaluation of telecommunication systems, notably when they have multiple access media. We consider two main methodologies:

- Deterministic performance analysis,
- Probabilistic performance analysis

In the deterministic analysis, the evaluation consists in identifying and quantifying the worst case scenario for an algorithm in a given context. For example to evaluate an end-to-end delay. Mathematically it consists into handling a (max,+) algebra. Since such algebra is not commutative, the complexity of the evaluation of an end-to-end delay frequently grows exponentially with the number of constraints. Therefore the main issue in the deterministic evaluation of performance is to find bounds easier to compute in order to have practical results in realistic situations.
In the probabilistic analysis of performance, one evaluate the behavior of an algorithm under a set of parameters that follows a stochastic model. For example traffic may be randomly generated, nodes may move randomly on a map. The pionneer works in this area come from Knuth (1973) who has systematized this branch. In the domain of telecommunication, the domain has started a significant rise with the appearance of the problematic of collision resolution in a multiple access medium. With the rise of wireless communication, new interesting problems have been investigated.
The analysis of algorithm can rely on analytical methodology which provides the better insight but is practical in very simplistic models. Simulation tools can be used to refine results in more complicated models. At the end of the line, we proceed with real life experiments. To simplify, experiments check the algorithms with 10 nodes in maximum, simulations with 100 nodes maximum, analytical tools with more 1,000 nodes, so that the full range of applicability of the algorithms is investigated.

3.2. Traffic and network architecture modeling

One needs good and realistic models of communication scenarios in order to provide pertinent performance evaluation of protocols. The models must assess the following key points:

- The architecture and topology: the way the nodes are structured within the network
- The mobility: the way the nodes move
- The dynamics: the way the nodes change status
- The traffic: the way the nodes communicate

For the architecture there are several scales. At the internet scale it is important to identify the patterns which dictate the node arrangement. For example the internet topology involves many power law distribution in node degree, link capacities, round trip delays. These parameters have a strong impact in the performance of the global network. At a smaller scale there is also the question how the nodes are connected in a wireless network. There is a significant difference between indoor and outdoor networks. The two kinds of networks differ on wave propagation. In indoor networks, the obstacles such as walls, furniture, etc, are the main source of signal attenuations. In outdoor networks the main source of signal attenuation is the distance to the emitter. This lead to very different models which vary between the random graph model for indoor networks to the unit graph model for outdoor networks.

The mobility model is very important for wireless network. The way nodes move may impact the performance of the network. For example it determines when the network splits in distinct connected components or when these components merge. With random graph models, the mobility model can be limited to the definition of a link status holding time. With unit disk model the mobility model will be defined according to random speed and direction during random times or random distances. There are some minor complications on the border of the map.
The node dynamic addresses the elements that change inside the node. For example its autonomy, its bandwidth requirement, the status of server, client, etc. Pair to pair networks involve a large class of users who frequently change status. In a mobile ad hoc network, nodes may change status just by entering or leaving the coverage area.
The traffic model is very most important. There are plenty of literature about trafic models which arose when Poisson models was shown not to be accurate for real traffics, on web or on local area networks. Natural traffic shows long range dependencies that do not exist in Poisson traffic. There are still strong issues about the origin of this long range dependencies which are debated, however they have a great impact on network performance since congestions are more frequent. The origin are either from the distribution of file sizes exchanged over the net, or from the protocols used to exchange them. One way to model the various size is to consider on/off sources. Every time a node is on it transfers a file of various size. The TCP protocol has also an impact since it keeps a memory on the network traffic. One way to describe it is to use an on/off model (a source sending packets in transmission windows) and to look at the superposition of these on/off sources.

3.3. Algorithm design, evaluation and implementation

The conception of algorithms is an important focus of the team. We specify algorithms in the perspective of achieving the best performance for communication. We also strive to embed those algorithms in protocols that involve the most legacy from existing technologies (Operating systems, internet, Wifi). Our aim with this respect is to allow code implementations for real life experiment or embedded simulation with existing network simulators. The algorithm specified by the project ranges from multiple access schemes, wireless ad hoc routing, to deployment of wireless sensor nodes as well as joint time slot and channel assignment in wireless networks. In any of these cases the design emphasize the notions of performance, robustness and flexibility. For example, a flooding technique in mobile ad hoc network should save bandwidth but should not stick too much close to optimal in order to be more reactive to frequent topology changes. Some telecommunication problems have NP hard optimal solution, and an implementable algorithm should be portable on very low power processing unit (e.g. sensors). Compromise have to be found and quantified with respect to nearly optimal solution.

3.4. Simulation of network algorithms and protocols

The performance of algorithms and protocols designed by the team have to be evaluated in various conditions: various configurations and various scenarii. The team uses different simulation tools. Historically, the first one was NS2 and some deployment algorithms are developed with NS2, taking advantage of its library and our previous works. We are now contributing to the development of NS3, enriching it with new modules (e.g. wireless medium access). For rapid simulation results and to validate design choices, we resort to Java home-made simulation tools (e.g. joint time slot and channel allocation).

INDES Project-Team

3. Research Program

3.1. Parallelism, concurrency, and distribution

Concurrency management is at the heart of diffuse programming. Since the execution platforms are highly heterogeneous, many different concurrency principles and models may be involved. Asynchronous concurrency is the basis of shared-memory process handling within multiprocessor or multicore computers, of direct or fifo-based message passing in distributed networks, and of fifo- or interrupt-based event handling in web-based human-machine interaction or sensor handling. Synchronous or quasi-synchronous concurrency is the basis of signal processing, of real-time control, and of safety-critical information acquisition and display. Interfacing existing devices based on these different concurrency principles within HOP or other diffuse programming languages will require better understanding of the underlying concurrency models and of the way they can nicely cooperate, a currently ill-resolved problem.

3.2. Web and functional programming

We are studying new paradigms for programming Web applications that rely on multi-tier functional programming [6]. We have created a Web programming environment named HOP. It relies on a single formalism for programming the server-side and the client-side of the applications as well as for configuring the execution engine.

HOP is a functional language based on the SCHEME programming language. That is, it is a strict functional language, fully polymorphic, supporting side effects, and dynamically type-checked. HOP is implemented as an extension of the BIGLOO compiler that we develop [7]. In the past, we have extensively studied static analyses (type systems and inference, abstract interpretations, as well as classical compiler optimizations) to improve the efficiency of compilation in both space and time.

3.3. Security of diffuse programs

The main goal of our security research is to provide scalable and rigorous language-based techniques that can be integrated into multi-tier compilers to enforce the security of diffuse programs. Research on language-based security has been carried on before in former Inria teams [2], [1]. In particular previous research has focused on controlling information flow to ensure confidentiality.

Typical language-based solutions to these problems are founded on static analysis, logics, provable cryptography, and compilers that generate correct code by construction [4]. Relying on the multi-tier programming language HOP that tames the complexity of writing and analysing secure diffuse applications, we are studying language-based solutions to prominent web security problems such as code injection and cross-site scripting, to name a few.

INFINE Team

3. Research Program

3.1. Online Social Networks (OSN)

Large-scale online social networks such as Twitter or FaceBook provide a powerful means of selecting information. They rely on "social filtering", whereby pieces of information are collectively evaluated and sorted by users. This gives rise to information cascades when one item reaches a large population after spreading much like an epidemics from user to user in a viral manner. Nevertheless, such OSNs expose their users to a large amount of content of no interest to them, a sign of poor "precision" according to the terminology of information retrieval. At the same time, many more relevant content items never reach those users most interested in them. In other words, OSNs also suffer from poor "recall" performance.

This leads to a first challenge: *what determines the optimal trade-off between precision and recall in OSNs? And what mechanisms should be deployed in order to approach such an optimal trade-off?* We intend to study this question at a theoretical level, by elaborating models and analyses of social filtering, and to validate the resulting hypotheses and designs through experimentation and processing of data traces. More specifically, we envision to reach this general objective by solving the following problems.

3.1.1. Community Detection

Identification of implicit communities of like-minded users and contact recommendation for helping users "rewire" the information network for better performance. Potential schemes may include variants of spectral clustering and belief propagation-style message passing. Limitations / relative merits of candidate schemes, their robustness to noise in the input data, will be investigated.

3.1.2. Incentivization

Design of incentive mechanisms to limit the impact of users' selfishness on system behavior: efficiency should be maintained even when users are gaming the system to try and increase their estimated expertise. By offering rewards to users on the basis of their involvement in filtering and propagation of content, one might encourage them to adjust their action and contribute to increase the overall efficiency of the OSN as a content access platform.

One promising direction will be to leverage the general class of Vickrey-Clarke-Groves incentive-compatible mechanisms of economic theory to design so-called marginal utility reward mechanisms for OSN users.

3.1.3. Social Recommendation and Privacy

So far we have only alluded to the potential benefits of OSNs in terms of better information access. We now turn to the risks they create. Privacy breaches constitute the greatest of these risks: OSN users disclose a wealth of personal information and thereby expose themselves to discrimination by potential employers, insurers, lenders, government agencies...Such privacy concerns are not specific to OSNs: internauts' online activity is discretely tracked by companies such as Bluekai, and subsequently monetized to advertisers seeking better ad targeting. While disclosure of personal data creates a privacy risk, on the other hand it fuels personalized services and thereby potentially benefits everyone.

One line of research will be to focus on the specific application scenario of content categorization, and to characterize analytically the trade-off between user privacy protection (captured by differential privacy), accuracy of content categorization, and sample complexity (measured in number of probed users).

3.2. Traffic and resource management

Despite the massive increases in transmission capacity of the last few years, one has every reason to believe that networks will remain durably congested, driven among other factors by the steadily increasing demand for video content, the proliferation of smart devices (i.e., smartphones or laptops with mobile data cards), and the forecasted additional traffic due to machine-to-machine (M2M) communications. Despite this rapid traffic growth, there is still a rather limited understanding of the features protocols have to support, the characteristics of the traffic being carried and the context where it is generated. There is thus a strong need for smart protocols that transport requested information at the cheapest possible cost on the network as well as provide good quality of service to network subscribers. One particularly new aspect of up-and-coming networks is that networks are now used to not only (i) access information, but also (ii) distributively process information, en-route.

We intend to study these issues at the theoretical and protocol design levels, by elaborating models and analysis of content demands and/or mobility of network subscribers. The resulting hypothesis and designs will be validated through experimentation, simulation, or data trace processing. It is also worth mentioning the provided solutions may bring benefits to different entities in the network: to content owners (if applied at the core of Internet) or to subscribers or network operators (if applied at the edge of the Internet).

3.2.1. At the Internet Core

One important optimization variable consists in content replication: users can access the closest replica of the content they are interested in. Thus the memory resource can be used to create more replicas and reduce the usage of the bandwidth resource. Another interesting arbitrage between resources arises because content is no longer static but rather dynamic. Here are two simple examples: i) a video could be encoded at several resolutions. There is then a choice between pre-recording all possible resolutions, or alternatively synthesizing a lower-resolution version on the fly from a higher resolution version when a request arises. ii) A user requests the result of a calculation, say the average temperature in a building; this can either be kept in memory, or recomputed each time such a query arises. Optimizing the joint use of all three resources, namely bandwidth, memory, computation, is a complex task. Content Delivery Network companies such as Akamai or Limelight have worked on the memory/bandwidth trade-off for some years, but as we will explain more can be done on this. On the other hand optimizing the memory/computation trade-off has received far less attention. We aim to characterize the best possible content replication strategies by leveraging fine-grained prediction of i) users' future requests, and ii) wireless channels' future bandwidth fluctuations. In the past these two determining inputs have only been considered at a coarse-grained, aggregate level. It is important to assess how much bandwidth saving can be had by conducting finer-grained prediction. We are developing light-weight protocols for conducting these predictions and automatically instantiating the corresponding optimal replication policies. We are also investigating generic protocols for automatically trading replication for computation, focusing initially on the above video transcoding scenario.

3.2.2. At the Internet Edge

Cellular and wireless data networks are increasingly relied upon to provide users with Internet access on devices such as smartphones, laptops or tablets. In particular, the proliferation of handheld devices equipped with multiple advanced capabilities (e.g., significant CPU and memory capacities, cameras, voice to text, text to voice, GPS, sensors, wireless communication) has catalyzed a fundamental change in the way people are connected, communicate, generate and exchange data. In this evolving network environment, users' social relations, opportunistic resource availability, and proximity between users' devices are significantly shaping the use and design of future networking protocols.

One consequence of these changes is that mobile data traffic has recently experienced a staggering growth in volume: Cisco has recently foreseen that the mobile data traffic will increase 18-fold within 2016, in front of a mere 9-fold increase in connection speeds. Hence, one can observe today that the inherently centralized and terminal-centric communication paradigm of currently deployed cellular networks cannot cope with the increased traffic demand generated by smartphone users. This mismatch is likely to last because (1) forecasted

mobile data traffic demand outgrows the capabilities of planned cellular technological advances such as 4G or LTE, and (2) there is strong skepticism about possible further improvements brought by 5G technology.

Congestion at the Internet's edge is thus here to stay. Solutions to this problem relates to: densify the infrastructure, opportunistically forward data among neighbors wireless devices, to offload data to alternate networks, or to bring content from the Internet closer to the subscribers. Our recent work on leveraging user mobility patterns, contact and inter-contact patterns, or content demand patterns constitute a starting point to these challenges. The projected increase of mobile data traffic demand pushes towards additional complementary offloading methods. Novel mechanisms are thus needed, which must fit both the new context that Internet users experience now, and their forecasted demands. In this realm, we will focus on new approaches leveraging ultra-distributed, user-centric approaches over IP.

3.3. Spontaneous Wireless Networks (SWN) and Internet of Things (IoT)

The unavailability of end-to-end connectivity in emergent wireless mobile networks is extremely disruptive for IP protocols. In fact, even in simpler cases of spontaneous wireless networks where end-to-end connectivity exists, such networks are still disruptive for the standard IP protocol stack, as many protocols rely on atomic link-local services (such as link-local multicast/broadcast), while these services are inherently unavailable in such networks due to their opportunistic, wireless multi hop nature. In this domain, we will aim to characterize the achievable performance in such IP-disruptive networks and to actively contribute to the design of new, deployable IP protocols that can tolerate these disruptions, while performing well enough compared to what is achievable and remaining interoperable with the rest of the Internet.

Spontaneous wireless networking is also a key aspect of the Internet of Things (IoT). The IoT is indeed expected to massively use this networking paradigm to gradually connect billions of new devices to the Internet, and drastically increase communication without human source or destination – to the point where the amount of such communications will dwarf communications involving humans. Large scale user environment automation require communication protocols optimized to efficiently leverage the heterogeneous and unreliable wireless vicinity (the scope of which may vary according to the application). In fact, extreme constraints in terms of cost, CPU, battery and memory capacities are typically experienced on a substantial fraction of IoT devices. We expect that such constraints will not vanish any time soon for two reasons. On one hand the progress made over the last decade concerning the cost/performance ratio for such small devices is quite disappointing. On the other hand, the ultimate goal of the IoT is ubiquitous Internet connectivity between devices as tiny as dust particles. These constraints actually require to redesign not only the network protocol stack running on these devices, but also the software platform powering these machines. In this context, we will aim at contributing to the design of novel network protocols and software platforms optimized to fit these constraints while remaining compatible with legacy Internet.

3.3.1. Design & Development of Open Experimental IoT Platforms

Based initially on "Demonstration abstract: Simply RIOT — Teaching and experimental research in the Internet of Things" Manufacturers announce on a regular basis the availability of novel tiny devices, most of them featuring network interfaces: the Internet of Things (IoT) is already here, from the hardware perspective, and it is expected in the near future that we will see a massive increase of the number of muti-purpose smart objects (from tiny sensors in industrial automation to devices like smart watches and tablets). Thus, one of the challenges is to be able to test architectures, protocols and applications, in realistic conditions and at large scale.

One necessity for research in this domain is to establish and improve IoT hardware platforms and testbeds, that integrate representative scenarios (such as Smart Energy, Home Automation etc.) and follow the evolution of technology, including radio technologies, and associated experimentation tools. For that, we plan to build upon the IoT-LAB federated testbeds, that we have participated in designing and deploying recently. We plan to further develop IoT-LAB with more heterogeneous, up-to-date IoT hardware and radios that will provide a usable and realistic experimentation environment. The goal is to provide a tool that enables testing an validation of upcoming software platforms and network stacks targeting concrete IoT deployments.

In parallel, on the software side, IoT hardware available so far made it uneasy for developers to build apps that run across heterogeneous hardware platforms. For instance Linux does not scale down to small, energy-constrained devices, while microcontroller-based OS alternatives were so far rudimentary and yield a steep learning curve and lengthy development life-cycles because they do not support standard programming and debugging tools. As a result, another necessity for research in this domain is to allow the emergence of it more powerful, unifying IOT software platforms, to bridge this gap. For that, we plan to build upon RIOT, a new open source software platform which provides a portable, Linux-like API for heterogeneous IoT hardware. We plan to continue to develop the systems and network stacks aspects of RIOT, within the open source developer community currently emerging around RIOT, which we co-founded together with Freie Universitaet Berlin. The key challenge is to improve usability and add functionalities, while maintaining architectural consistency and a small enough memory footprint. The goal is to provide an IoT software platform that can be used like Linux is used for less constrained machines, both (i) in the context of research and/or teaching, as well as (ii) in industrial contexts. Of course, we plan to use it ourselves for our own experimental research activities in the domain of IoT e.g., as an API to implement novel network protocols running on IoT hardware, to be tested and validated on IoT-LAB testbeds.

3.3.2. Design & Standardization of Architectures and Efficient Protocols for Internet of Things

As described before, and by definition, the Internet of Things will integrate not only a massive number of homogeneous devices (e.g., networks of wireless sensors), but also heterogeneous devices using various communication technologies. Most devices will be very constrained resources (memory resources, computational resources, energy). Communicating with (and amongst) such devices is a key challenge that we will focus on. The ability to communicate efficiently, to communicate reliably, or even just to be able to communicate at all, is non-trivial in many IoT scenarios: in this respect, we intend to develop innovative protocols, while following and contributing to standardization in this area. We will focus and base most of our work on standards developed in the context of the IETF, in working groups such as 6lo, CORE, LWIG etc., as well as IRTF research groups such as NWCRG on network coding and ICNRG on Information Centric Networking. We note however that this task goes far beyond protocol design: recently, radical rearchitecturing of the networks with new paradigms such as Information Centric Networking, ICN, (or even in wired networks, software-defined networks), have opened exciting new avenues. One of our direction of research will be to explore these content-centric approaches, and other novel architectures, in the context of IoT.

KerData Project-Team

3. Research Program

3.1. Our goals and methodology

Data-intensive applications demonstrate common requirements with respect to the need for data storage and I/O processing. These requirements lead to several core challenges discussed below.

Challenges related to cloud storage. In the area of cloud data management, a significant milestone is the emergence of the Map-Reduce [31] parallel programming paradigm, currently used on most cloud platforms, following the trend set up by Amazon [27]. At the core of Map-Reduce frameworks lies a key component, which must meet a series of specific requirements that have not fully been met yet by existing solutions: the ability to provide efficient *fine-grain access* to the files, while sustaining a *high throughput* in spite of *heavy access concurrency*. Additionally, as thousands of clients simultaneously access shared data, it is critical to preserve *fault-tolerance* and *security* requirements.

Challenges related to data-intensive HPC applications. The requirements exhibited by climate simulations specifically highlight a major, more general research topic. They have been clearly identified by international panels of experts like IESP [30], EESI [28], ETP4HPC [29] in the context of HPC simulations running on post-Petascale supercomputers. A jump of one order of magnitude in the size of numerical simulations is required to address some of the fundamental questions in several communities such as climate modeling, solid earth sciences or astrophysics. In this context, the lack of data-intensive infrastructures and methodologies to analyze huge simulations is a growing limiting factor. The challenge is to find new ways to store and analyze massive outputs of data during and after the simulation without impacting the overall performance.

The overall goal of the KerData project-team is to bring a substantial contribution to the effort of the research community to address the above challenges. KerData aims to design and implement distributed algorithms for scalable data storage and input/output management for efficient large-scale data processing. We target two main execution infrastructures: cloud platforms and post-Petascale HPC supercomputers. Additionally, we are also looking at other kinds of infrastructures, e.g. hybrid platforms combining enterprise desktop grids extended to cloud platforms. Our collaboration porfolio includes international teams that are active in this area both in Academia (e.g., Argonne National Lab, University of Illinois at Urbana-Champaign, Barcelona Supercomputing Centre) and Industry (Microsoft, IBM).

The highly experimental nature of our research validation methodology should be stressed. Our approach relies on building prototypes and on validating them at a large scale on real testbeds and experimental platforms. We strongly rely on the Grid'5000 platform. Moreover, thanks to our projects and partnerships, we have access to reference software and physical infrastructures in the cloud area (Microsoft Azure, Amazon clouds, Nimbus clouds); in the post-Petascale HPC area we have access to the Jaguar and Kraken supercomputers (ranked 3rd and 11th respectively in the Top 500 supercomputer list) and to the Blue Waters supercomputer. This provides us with excellent opportunities to validate our results on advanced realistic platforms.

Moreover, the consortiums of our current projects include application partners in the areas of Bio-Chemistry, Neurology and Genetics, and Climate Simulations. This is an additional asset, it enables us to take into account application requirements in the early design phase of our solutions, and to validate those solutions with real applications. We intend to continue increasing our collaborations with application communities, as we believe that this a key to perform effective research with a high impact.

3.2. Our research agenda

Three typical application scenarios will be described in detail in the next section:

- Joint genetic and neuroimaging data analysis on Azure clouds;
- Structural protein analysis on Nimbus clouds;
- I/O intensive climate simulations for the Blue Waters post-Petascale machine.

They illustrate the above challenges in some specific ways. They all exhibit a common scheme: massively concurrent processes which access massive data at a fine granularity, where data is shared and distributed at a large scale. To address the aforementioned challenges efficiently, we have started to work out an approach called BlobSeer, which stands today at the center of our research efforts. This approach relies on the design and implementation of *scalable* distributed algorithms for data storage and access. They combine advanced techniques for decentralized metadata and data management, with versioning-based concurrency control to optimize the performance of applications under heavy access concurrency.

Preliminary experiments with our BlobSeer BLOB management system within today's cloud software infrastructures proved very promising. Recently, we used the BlobSeer approach as a starting point to address two usage scenarios in more detail, which led to two more specific approaches: 1) Pyramid [35] (which borrows many concepts from BlobSeer), with a specific focus on array-oriented storage; and 2) Damaris (totally independent of BlobSeer), which exploits multicore parallelism in post-Petascale supercomputers. All these directions are described below.

Our short- and medium-term research plan is devoted to storage challenges in two main contexts: clouds and post-Petascale HPC architectures. Consequently, our research plan is split in two main themes, which correspond to their respective challenges. For each of those themes, we have initiated several actions through collaborative projects coordinated by KerData, which define our agenda for the next 4 years.

Based on very promising results demonstrated by BlobSeer in preliminary experiments [34], we have initiated several collaborative projects in the area of cloud data management, e.g., the MapReduce ANR project, the A-Brain Microsoft-Inria project, the Z-CloudFlow Microsoft-Inria project. Such frameworks are for us concrete and efficient means to work in close connection with strong partners already well positioned in the area of cloud computing research. Thanks to these projects, we have already started to enjoy a visible scientific positioning at the international level.

The particularly active Data@Exascale Associate Team creates the framework for an enlarged research activity involving a large number of young researchers and students. It serves as a basis for extended research activities based on our approaches, carried out beyond the frontiers of our team. In the HPC area, our presence in the research activities of the Joint UIUC-Inria Lab for Petascale Computing (JLPC) at Urbana-Champaign is a very exciting opportunity that we have started to leverage. It facilitates high-quality collaborations and access to some of the most powerful supercomputers, an important asset which already helped us produce and transfer some results, as described in Section 6.5 .

MADYNES Project-Team

3. Research Program

3.1. Evolutionary needs in network and service management

The foundation of the MADYNES research activity is the ever increasing need for automated monitoring and control within networked environments. This need is mainly due to the increasing dependency of both people and goods towards communication infrastructures as well as the growing demand towards services of higher quality. Because of its strategic importance and crucial requirements for interoperability, the management models were constructed in the context of strong standardization activities by many different organizations over the last 15 years. This has led to the design of most of the paradigms used in today's deployed approaches. These paradigms are the Manager/Agent interaction model, the Information Model paradigm and its container, together with a naming infrastructure called the Management Information Base. In addition to this structure, five functional areas known under Fault, Configuration, Accounting, Performance and Security are associated to these standards.

While these models were well suited for the specific application domains for which they were designed (telecommunication networks or dedicated protocol stacks), they all show the same limits. Especially they are unable:

1. to deal with any form of dynamicity in the managed environment,
2. to master the complexity, the operating mode and the heterogeneity of the emerging services,
3. to scale to new networks and service environments.

These three limits are observed in all five functional areas of the management domain (fault, configuration, accounting, performance and security) and represent the major challenges when it comes to enable effective automated management and control of devices, networks and services in the next decade.

MADYNES addresses these challenges by focusing on the design of management models that rely on inherently dynamic and evolving environments. The project is centered around two core activities. These activities are, as mentioned in the previous section, the design of an autonomous management framework and its application to three of the standard functional areas namely security, configuration and performance.

3.2. Autonomous management

3.2.1. Models and methods for a self-management plane

Self organization and automation are fundamental requirements within the management plane in today's dynamic environments. It is necessary to automate the management processes and enable management frameworks to operate in time sensitive evolving networks and service environments. The automation of the organization of devices, software components, networks and services is investigated in many research projects and has already led to several solution proposals. While these proposals are successful at several layers, like IP auto-configuration or service discovery and binding facilities, they did not enhance the management plane at all. For example, while self-configuration of IP devices is commonplace, no solution exists that provides strong support to the management plane to configure itself (e.g. finding the manager to which an agent has to send traps or organizing the access control based on locality or any other context information). So, this area represents a major challenge in extending current management approaches so that they become self-organized.

Our approach is bottom-up and consists in identifying those parameters and framework elements (manager data, information model sharing, agent parameters, protocol settings, ...) that need dynamic configuration and self-organization (like the address of a trap sink). For these parameters and their instantiation in various management frameworks (SNMP, Netconf, WBEM, ...), we investigate and elaborate novel approaches enabling fully automated setup and operation in the management plane.

3.2.2. Design and evaluation of P2P-based management architectures

Over the last years, several models have emerged and gained wide acceptance in the networking and service world. Among them, the overlay networks together with the P2P paradigms appear to be very promising. Since they rely mainly on fully decentralized models, they offer excellent fault tolerance and have a real potential to achieve high scalability. Mainly deployed in the content delivery and the cooperation and distributed computation disciplines, they seem to offer all features required by a management framework that needs to operate in a dynamic world. This potential however needs an in depth investigation because these models have also many characteristics that are unusual in management (e.g. a fast and uncontrolled evolution of the topology or the existence of a distributed trust relationship framework rather than a standard centralized security framework).

Our approach envisions how a complete redesign of a management framework is done given the characteristics of the underlying P2P and overlay services. Among the topics of interest we study the concept of management information and operations routing within a management overlay as well as the distribution of management functions in a multi-manager/agent P2P environment. The functional areas targeted in our approach by the P2P model are network and service configuration and distributed monitoring. The models are to be evaluated against highly dynamic frameworks such as ad-hoc environments (network or application level) and mobile devices.

3.2.3. Integration of management information

Representation, specification and integration of management information models form a foundation for network and service management and remains an open research domain. The design and specification of new models is mainly driven by the appearance of new protocols, services and usage patterns. These need to be managed and exposed through well designed management information models. Integration activities are driven by the multiplication of various management approaches. To enable automated management, these approaches need to inter-operate which is not the case today.

The MADYNES approach to this problem of modeling and representation of management information aims at:

1. enabling application developers to establish their management interface in the same workspace, with the same notations and concepts as the ones used to develop their application,
2. fostering the use of standard models (at least the structure and semantics of well defined models),
3. designing a naming structure that allows the routing of management information in an overlay management plane, and
4. evaluating new approaches for management information integration especially based on management ontologies and semantic information models.

3.2.4. Modeling and benchmarking of dynamic networks

The impact of a management approach on the efficiency of the managed service is highly dependent on three factors:

- the distribution of the considered service and their associated management tasks,
- the management patterns used (e.g. monitoring frequency, granularity of the management information considered),
- the cost in terms of resources these considered functions have on the managed element (e.g. method call overhead, management memory footprint).

MADYNES addresses this problem from multiple viewpoints: communication patterns, processing and memory resources consumption. Our goal is to provide management patterns combining optimized management technologies so as to optimize the resources consumed by the management activity imposed by the operating environment while ensuring its efficiency in large dynamic networks.

3.3. Functional areas

3.3.1. Security management

Securing the management plane is vital. While several proposals are already integrated in the existing management frameworks, they are rarely used. This is due to the fact that these approaches are completely detached from the enterprise security framework. As a consequence, the management framework is "managed" separately with different models; this represents a huge overhead. Moreover the current approaches to security in the management plane are not inter-operable at all, multiplying the operational costs in a heterogeneous management framework.

The primary goal of the research in this activity is the design and the validation of a security framework for the management plane that will be open and capable to integrate the security services provided in today's management architectures. Management security interoperability is of major importance in this activity.

Our activity in this area aims at designing a generic security model in the context of multi-party / multi-technology management interactions. Therefore, we develop research on the following directions:

1. Abstraction of the various access control mechanisms that exist in today's management frameworks. We are particularly interested in extending these models so that they support event-driven management, which is not the case for most of them today.
2. Extension of policy and trust models to ease and to ensure coordination among managers towards one agent or a subset of the management tree. Provisional policies are of great interest to us in this context.
3. Evaluation of the adequacy of key distribution architectures to the needs of the management plane as well as selecting reputation models to be used in the management of highly dynamic environments (e.g. multicast groups, ad-hoc networks).

A strong requirement towards the future generic model is that it needs to be instantiated (with potential restrictions) into standard management platforms like SNMP, WBEM or Netconf and to allow interoperability in environments where these approaches coexist and even cooperate. A typical example of this is the security of an integration agent which is located in two management worlds.

Since 2006 we have also started an activity on security assessment. The objective is to investigate new methods and models for validating the security of large scale dynamic networks and services. The first targeted service is VoIP.

3.3.2. Configuration: automation of service configuration and provisioning

Configuration covers many processes which are all important to enable dynamic networks. Within our research activity, we focus on the operation of tuning the parameters of a service in an automated way. This is done together with the activation topics of configuration management and the monitoring information collected from the underlying infrastructure. Some approaches exist today to automate part of the configuration process (download of a configuration file at boot time within a router, on demand code deployment in service platforms). While these approaches are interesting they all suffer from the same limits, namely:

1. they rely on specific service life cycle models,
2. they use proprietary interfaces and protocols.

These two basic limits have high impacts on service dynamics in a heterogeneous environment.

We follow two research directions in the topic of configuration management. The first one aims at establishing an abstract life-cycle model for either a service, a device or a network configuration and to associate with this model a generic command and programming interface. This is done in a way similar to what is proposed in the area of call control in initiatives such as Parlay or OSA.

In addition to the investigation of the life-cycle model, we work on technology support for distributing and exchanging configuration management information. Especially, we investigate policy-driven approaches for representing configurations and constraints while we study XML-based protocols for coordinating distribution and synchronization. Off and online validation of configuration data is also part of this effort.

3.3.3. Performance and availability monitoring

Performance management is one of the most important and deployed management function. It is crucial for any service which is bound to an agreement about the expected delivery level. Performance management needs models, metrics, associated instrumentation, data collection and aggregation infrastructures and advanced data analysis algorithms.

Today, a programmable approach for end-to-end service performance measurement in a client server environment exists. This approach, called Application Response Measurement (ARM) defines a model including an abstract definition of a unit of work and related performance records; it offers an API to application developers which allows easy integration of measurement within their distributed application. While this approach is interesting, it is only a first step toward the automation of performance management.

We are investigating two specific aspects. First we are working on the coupling and possible automation of performance measurement models with the upper service level agreement and specification levels. Second we are working on the mapping of these high level requirements to the lower level of instrumentation and actual data collection processes available in the network. More specifically we are interested in providing automated mapping of service level parameters to monitoring and measurement capabilities. We also envision automated deployment and/or activation of performance measurement sensors based on the mapped parameters. This activity also incorporates self-instrumentation (and when possible on the fly instrumentation) of software components for performance monitoring purpose.

MAESTRO Project-Team

3. Research Program

3.1. Research Directions

MAESTRO's research directions belong to five main themes motivated by direct applications: network science, wireless networks, network engineering games, green networking and smart grids, content-oriented systems. These directions are very connected: network engineering games find applications in many networking fields, from wireless protocols to applications such as social networks. Green IT studies are often concerned with wireless networks, etc. The study of these applications often raises questions of methodological nature, less close to direct applications; these advances are reported in a separate section.

3.1.1. Network Science

MAESTRO contributes to this new fast growing research subject. "Network Science" or "Complex Network Analysis" aims at understanding the structural properties and the dynamics of a variety of large-scale networks in telecommunications (e.g. the graph of autonomous systems, the Web graph), social science (e.g. community of interest, advertisement, reputation, recommendation systems), bibliometrics (e.g. citations, co-authors), biology (e.g. spread of an epidemic, protein-protein interactions), and physics. It has been observed that the complex networks encountered in these areas share common properties such as power law degree distribution, small average distances, community structure, etc. It also appears that many general questions/applications (e.g. community detection, epidemic spreading, search, anomaly detection) are common in various disciplines which study networks. In particular, we aim at understanding the evolution of complex networks with the help of game theoretical tools in connection with Network Engineering Games, as described below. We design efficient tools for measuring specific properties of large scale complex networks and their dynamics. More specifically, we work on the problem of distributed optimization in large networks where nodes cooperatively solve an optimization problem relying only on local information exchange.

3.1.2. Wireless Networks

The amazing technological advances in wireless devices has led networks to become heterogeneous and very complex. Many research groups worldwide investigate performance evaluation of wireless technologies. MAESTRO's specificity relies on the use of a large variety of analytic tools from applied probability, control theory and distributed optimization to study and improve wireless network functionalities.

3.1.3. Network Engineering Games

The foundations of *Network Engineering Games* are currently being laid. These are games arising in telecommunications engineering at all the networking layers. This includes considerations from information and communications theory for dealing with the physical and link layers, along with cross layer approaches. MAESTRO's focus is on three areas: *routing games*, *evolutionary games* and *epidemic games*. In routing games we progress on the theory for costs that are not additive over links (such as packet losses or call blocking probabilities). We pursue our research in the stochastic extension of evolutionary game theory, namely the "anonymous sequential games" in which we study the total expected costs and the average cost. Within epidemic games we study epidemics that compete against each other. We apply this to social networks, considering in particular the coupling between various social networks (e.g. propagation strategies that combine Twitter, FaceBook and other social networks).

3.1.4. Green Networking and Smart Grids

The ICT (Information and Communications Technology) sector is becoming one of the main energy consumers worldwide. There is awareness that networks should have a reduced environmental footprint. Our objective is to have a systematically "green" approach when solving optimization problems. The energy cost and the environmental impact should be considered in optimization functions along with traditional performance metrics such as throughput, fairness or delay. We aim at contributing to the design and the analysis of future green networks, in particular those using renewable energy.

Researchers envision that future electricity distribution network will be “smart”, with a large number of small generators (due to an extensive use of renewable energies) and of consumer devices able to adapt their energy needs to a time-varying offer. Generators and devices will be able to locally communicate through the electrical grid itself (or more traditional communication networks), in order to optimize production, transport and use of the energy. This is definitely a new application scenario for MAESTRO, to which we hope to be able to contribute with our expertise on analytic models and performance evaluation.

3.1.5. Content-Oriented Systems

We generally study problems related with the placement and the retrieval of data in communication networks.

We are particularly interested in In-network caching, a widely adopted technique to provide an efficient access to data or resources on a world-wide deployed system while ensuring scalability and availability. For instance, caches are integral components of the Domain Name System, the World Wide Web, Content Distribution Networks, or the recently proposed Information-Centric Network (ICN) architectures. We analyze network of caches, study their optimal placement in the network and optimize data placement in caches/servers.

We also study other aspects related to replication and placement of data: how much to replicate it and on which servers to place it? Finally, we study optimal ways of retrieving the data through prefetching.

3.1.6. Advances in Methodological Tools

MAESTRO has a methodological activity that aims at advancing the state of the art in the methodological tools used for the general performance evaluation and control of systems. We contribute to such fields as perturbation analysis, Markov processes, queueing theory, control theory and game theory. Another objective is to enhance our activity on general-purpose modeling algorithms and software for controlled and uncontrolled stochastic systems.

3.2. Scientific Foundations

The main mathematical tools and formalisms used in MAESTRO include:

- theory of stochastic processes: Markov process, renewal process, branching process, point process, Palm measure, large deviations, mean-field approximation, fluid approximation;
- theory of dynamical discrete-event systems: queues, pathwise and stochastic comparisons, random matrix theory;
- theory of control and scheduling: dynamic programming, Markov decision process, game theory, deterministic and stochastic scheduling; stochastic approximation algorithms;
- theory of singular perturbations.

MESCAL Project-Team

3. Research Program

3.1. Large System Modeling and Analysis

Participants: Nicolas Gast, Bruno Gaujal, Arnaud Legrand, Panayotis Mertikopoulos, Florence Perronnin, Olivier Richard, Jean-Marc Vincent.

Markov chains, Queuing networks, Mean field approximation, Simulation, Performance evaluation, Discrete event dynamic systems.

3.1.1. Simulation of distributed systems

Since the advent of distributed computer systems, an active field of research has been the investigation of *scheduling* strategies for parallel applications. The common approach is to employ scheduling heuristics that approximate an optimal schedule. Unfortunately, it is often impossible to obtain analytical results to compare the efficiency of these heuristics. One possibility is to conduct large numbers of back-to-back experiments on real platforms. While this is possible on tightly-coupled platforms, it is unfeasible on modern distributed platforms (i.e., grids or peer-to-peer environments) as it is labor-intensive and does not enable repeatable results. The solution is to resort to *simulations*.

3.1.1.1. Flow Simulations

To make simulations of large systems efficient and trustful, we have used flow simulations (where streams of packets are abstracted into flows). SimGrid is a simulation platform that specifically targets the simulation of large distributed systems (grids, clusters, peer-to-peer systems, volunteer computing systems, clouds) from the perspective of applications. It enables to obtain repeatable results and to explore wide ranges of platform and application scenarios.

3.1.1.2. Perfect Simulation

Using a constructive representation of a Markovian queuing network based on events (often called GSMPs), we have designed perfect simulation algorithms computing samples distributed according to the stationary distribution of the Markov process with no bias. The tools based on our algorithms (ψ) can sample the stationary measure of Markov processes using directly the queuing network description. Some monotone networks with up to 10^{50} states can be handled within minutes over a regular PC.

3.1.2. Fluid models and mean field limits

When the size of systems grows very large, one may use asymptotic techniques to get a faithful estimate of their behavior. One such tool is mean field analysis and fluid limits, that can be used at a modeling and simulation level. Proving that large discrete dynamic systems can be approximated by continuous dynamics uses the theory of stochastic approximation pioneered by Michel Benaïm or population dynamics introduced by Thomas Kurtz and others. We have extended the stochastic approximation approach to take into account discontinuities in the dynamics as well as to tackle optimization issues.

Recent applications include call centers and peer to peer systems, where the mean field approach helps to get a better understanding of the behavior of the system and to solve several optimization problems. Another application concerns task brokering in desktop grids taking into account statistical features of tasks as well as of the availability of the processors. Mean field has also been applied to the performance evaluation of work stealing in large systems and to model central/local controllers as well as knitting systems.

3.1.3. Game Theory

Resources in large-scale distributed platforms (grid computing platforms, enterprise networks, peer-to-peer systems) are shared by a number of users having conflicting interests who are thus prone to act selfishly. A natural framework for studying such non-cooperative individual decision-making is game theory. In particular, game theory models the decentralized nature of decision-making.

It is well known that such non-cooperative behaviors can lead to important inefficiencies and unfairness. In other words, individual optimizations often result in global resource waste. In the context of game theory, a situation in which all users selfishly optimize their own utility is known as a *Nash equilibrium* or *Wardrop equilibrium*. In such equilibria, no user has interest in unilaterally deviating from its strategy. Such policies are thus very natural to seek in fully distributed systems and have some stability properties. However, a possible consequence is the *Braess paradox* in which the increase of resource happens at the expense of *every* user. This is why, the study of the occurrence and degree of such inefficiency is of crucial interest. Up until now, little is known about general conditions for optimality or degree of efficiency of these equilibria, in a general setting.

Many techniques have been developed to enforce some form of collaboration and improve these equilibria. In this context, it is generally prohibitive to take joint decisions so that a global optimization cannot be achieved. A possible option relies on the establishment of virtual prices, also called *shadow prices* in congestion networks. These prices ensure a rational use of resources.

Once the payoffs are fixed (using shadow prices or not), the main question is to design algorithms that allow the players to learn Nash equilibria in a distributed way, while being robust to noise and information delay as well as fast enough to outrate changing conditions of the environment.

3.2. Management of Large Architectures

Participants: Nicolas Gast, Arnaud Legrand, Olivier Richard.

Administration, Deployment, Peer-to-peer, Clusters, Grids, Clouds, Job scheduler

3.2.1. Instrumentation, analysis and prediction tools

To understand complex distributed systems, one has to provide reliable measurements together with accurate models before applying this understanding to improve system design.

Our approach for instrumentation of distributed systems (embedded systems as well as multi-core machines or distributed systems) relies on quality of service criteria. In particular, we focus on non-obtrusiveness and experimental reproducibility.

Our approach for analysis is to use statistical methods with experimental data of real systems to understand their normal or abnormal behavior. With that approach we are able to predict availability of very large systems (with more than 100,000 nodes), to design cost-aware resource management (based on mathematical modeling and performance evaluation of target architectures), and to propose several scheduling policies tailored for unreliable and shared resources.

3.2.2. Fairness in large-scale distributed systems

Large-scale distributed platforms (grid computing platforms, enterprise networks, peer-to-peer systems) result from the collaboration of many people. Thus, the scaling evolution we are facing is not only dealing with the amount of data and the number of computers but also with the number of users and the diversity of their behavior. In a high-performance computing framework, the rationale behind this joining of forces is that most users need a larger amount of resources than what they have on their own. Some only need these resources for a limited amount of time. On the opposite some others need as many resources as possible but do not have particular deadlines. Some may have mainly tightly-coupled applications while some others may have mostly embarrassingly parallel applications. The variety of user profiles makes resources sharing a challenge. However resources have to be *fairly* shared between users, otherwise users will leave the group and join another one. Large-scale systems therefore have a real need for fairness and this notion is missing from classical scheduling models.

3.2.3. Tools to operate clusters

The MESCAL project-team studies and develops a set of tools designed to help the installation and the use of a cluster of PCs. The first version had been developed for the Icluster1 platform exploitation. The main tools are a scalable tool for cloning nodes (KA-DEPLOY) and a parallel launcher based on the TAKTUK project (now developed by the MOAIS project-team). Many interesting issues have been raised by the use of the first versions among which we can mention environment deployment, robustness and batch scheduler integration. A second generation of these tools is thus under development to meet these requirements.

KA-DEPLOY has been retained as the primary deployment tool for the experimental national grid Grid'5000.

3.2.4. Simple and scalable batch scheduler for clusters and grids

Most known batch schedulers (PBS, LSF, Condor, ...) are built in a monolithic way, with the purpose of fulfilling most of the exploitation needs. This results in systems of high software complexity (150,000 lines of code for OpenPBS), offering a growing number of functions that are, most of the time, not used. In such a context, it becomes hard to control both the robustness and the scalability of the whole system.

OAR is an attempt to address these issues. Firstly, OAR is written in a very high level language (Perl) and makes intensive use of high level tools (MySql and TAKTUK), thereby resulting in a concise code (around 5000 lines of code) easy to maintain and extend. This small code as well as the choice of widespread tools (MySql) are essential elements that ensure a strong robustness of the system. Secondly, OAR makes use of SQL queries to perform most of its job management tasks thereby getting advantage of the strong scalability of most database management tools. Such scalability is further improved in OAR by making use of TAKTUK to manage nodes themselves.

3.3. Migration resilience; Large scale data management

Participant: Yves Denneulin.

Fault tolerance, migration, distributed algorithms.

Most propositions to improve reliability address only a given application or service. This may be due to the fact that until clusters and intranet architectures arose, it was obvious that client and server nodes were independent. This is not the case in parallel scientific computing where a fault on a node can lead to a data loss on thousands of other nodes. MESCAL's work on this topic is based on the idea that each process in a parallel application will be executed by a group of nodes instead of a single node: when the node in charge of a process fails, another in the same group can replace it in a transparent way for the application.

There are two main problems to be solved in order to achieve this objective. The first one is the ability to migrate processes of a parallel, and thus communicating, application without enforcing modifications. The second one is the ability to maintain a group structure in a completely distributed way. They both rely on a close interaction with the underlying operating systems and networks, since processes can be migrated in the middle of a communication. This can only be done by knowing how to save and replay later all ongoing communications, independently of the communication pattern. Freezing a process to restore it on another node is also an operation that requires collaboration of the operating system and a good knowledge of its internals. The other main problem (keeping a group structure) belongs to the distributed algorithms domain and is of a higher level nature.

MIMOVE Team

3. Research Program

3.1. Introduction

MiMove targets research enabling next-generation mobile distributed systems, from their conception and design to their runtime support. These systems are challenged by their own success and consequent massive growth, as well as by the present and future, fast evolving, global networking and computing environment. This context is well-captured by the Future Internet vision, whose mobile constituents are becoming the norm rather than the exception. MiMove's research topics relate to a number of scientific domains with intensive ongoing research, such as ubiquitous computing, self-adaptive systems, wireless sensor networks, participatory sensing and social networks. In the following, we discuss related state-of-the-art research – in particular work focusing on middleware for mobile systems – and we identify the open research challenges that drive our work.

3.2. Emergent mobile distributed systems

Emergent mobile distributed systems promise to provide solutions to the complexity of the current and future computing and networking environments as well as to the ever higher demand for ubiquitous mobile applications, in particular being a response to the volatile and evolving nature of both the former and the latter. Hence, such systems have gained growing interest in the research literature. Notably, research communities have been formed around *self-adaptive systems* and *autonomic systems*, for which various overlapping definitions exist [84]. Self-adaptive systems are systems that are able to adapt themselves to uncertain execution environments, while autonomic systems have been defined as having one or more characteristics known as *self-** properties, including self-configuring, self-healing, self-optimizing and self-protecting [66]. Self-adaptive or autonomic systems typically include an adaptation loop comprising *modeling*, *monitoring*, *analyzing*, *deciding* and *enactment* processes. The adaptation loop provides feedback about changes in the system and its environment to the system itself, which adjusts itself in response. Current research on emergent distributed systems, including mobile ones, addresses all the dimensions of the adaptation loop [44], [39], [74], [95].

In our previous work, we introduced the paradigm of *emergent middleware*, which enables networked systems with heterogeneous behaviors to coordinate through adequate interaction protocols that emerge in an automated way [62], [41], [40]. A key point of that work is the combined study of the application- and middleware-layer behaviors, while current efforts in the literature tend to look only at one layer, either the application [60] or the middleware [34], [61], and take the other for granted (i.e., homogeneous, allowing direct coordination). Furthermore, the uncertainty of the computing and networking environments that is intrinsic to emergent mobile distributed systems [53] calls for taking into account also the underlying network and computational resources in a cross-layer fashion. In another line of work, we studied cross-integration of heterogeneous interaction paradigms at the middleware layer (message passing versus event-based and data sharing), where we investigate functional and QoS semantics of paradigms across their interconnections [55], [65]. Our focus there is to grasp the relation between individual and end-to-end semantics when bridging heterogeneous interaction protocols. In contrast, existing research efforts typically focus on emergent or evolving properties in homogeneous settings [54]. Last but not least, integrating heterogeneous mobile distributed systems into emergent compositions raises the question of dependability. More specifically, the overall correctness of the composition with respect to the individual requirements of the constituent systems can be particularly hard to ensure due to their heterogeneity. Again, current approaches typically deal with homogeneous constraints for dependability [51], [97], [52] with few exceptions [50].

As evident from the above, there is considerable interest and intensive research on emergent mobile distributed systems, while at the same time there are key research questions that remain open despite initial relevant work, including ours, which are summarized in the following:

- How to effectively deal with the combined impact on emergent properties of the different functional layers of mobile distributed systems (e.g., [62], [41], [40], [81])?
- How to perceive and model emergent properties in space and in time across volatile compositions of heterogeneous mobile distributed systems (e.g., [55], [65])?
- How to produce dependable emergent mobile distributed systems, i.e., systems that correctly meet their requirements, despite uncertainty in their emergence and execution exacerbated by heterogeneity (e.g., [50])?

3.3. Large-scale mobile sensing and actuation

In the past decade, the increasingly low cost of MEMS [0] devices and low-power microprocessors has led to a significant amount of research into mobile sensing and actuation. The results of this are now reaching the general public, going beyond the largely static use of sensors in scenarios such as agriculture and waste-water management, into increasingly *mobile* systems. These include sensor-equipped smartphones and personal wearable devices focused on the idea of a "quantified self", gathering data about a user's daily habits in order to enable them to improve their well-being. However, in spite of significant advances, the key challenges of these systems arise from largely the same attributes as those of early envisioned mobile systems, introduced in [88] and re-iterated in [87]: relative resource-poverty in terms of computation and communication, variable and unreliable connectivity, and limitations imposed by a finite energy source. These remain true even though modern mobile devices are significantly more powerful compared to their ancestors; the work we expect them to do has increased, and the computation and storage abilities available through fixed infrastructure such as the cloud are larger by order of magnitudes than any single mobile device. The design of algorithms and protocols to efficiently coordinate the sensing, processing, and actuation capabilities of the large number of mobile devices in future systems is a core area of MiMove's research.

Precisely, the focus of MiMove's research interests lies mostly in the systems resulting from the increased popularity of sensor-equipped smart devices that are carried by people, which has led to the promising field of *mobile phone sensing* or *mobile crowd-sensing* [71], [67]. The paradigm is powerful, as it allows overcoming the inherent limitation of traditional sensing techniques that require the deployment of dedicated fixed sensors (e.g., see work on noise mapping using the microphones in users' telephones [82]). Specifically, we are interested in the challenges below, noting that initial work to address them already exists, including that by team members:

- How to efficiently manage the large scale that will come to the fore when millions, even billions of devices will need to be managed and queried simultaneously (e.g., [93], [57])?
- How to efficiently coordinate the available devices, including resource-poor mobile devices and the more-capable cloud infrastructure (e.g., [80], [48], [86], [77])?
- How to guarantee dependability in a mobile computing environment (e.g., [47], [92], [43])?
- How to ensure that the overhead of sensing does not lead to a degraded performance for the user (e.g., [69], [48])?

3.4. Mobile social crowd-sensing

Mobile crowd-sensing as introduced in Section 3.3 is further undergoing a transformation due to the widespread adoption of social networking. The resulting mobile *social* crowd-sensing may be qualified as "*people-centric sensing*" and roughly subdivides into two categories [70]: i) *participatory sensing*, and ii) *opportunistic sensing*. Participatory sensing entails direct involvement of humans controlling the mobile devices, while opportunistic sensing requires the mobile device itself to determine whether or not to perform

[0] Micro-Electro-Mechanical Systems.

the sensing task. Orthogonally to the above categorization, mobile sensing can be [67]: i) *personal sensing*, mostly to monitor a person's context and well-being; ii) *social sensing*, where updates are about the social and emotional statuses of individuals; or iii) *urban (public) sensing*, where public data is generated by the public and for the public to exploit. Personal sensing is aimed towards personal monitoring and involves one or just a few devices in direct relationship with their custodian. For instance, SoundSense [75] is a system that enables each person's mobile device to learn the types of sounds the owner encounters through unsupervised learning. Another application example relates to the sensing-based detection of the users' transportation mode by using their smartphones [59]. In social sensing, the mobile device or its owner decides what social information to share about the owner or the owner's environment, with an individual or group of friends [67], [49], [64], [35], [79]. Social sensing is mostly participatory. Therefore, it is the custodian of the device who determines when and where data should be generated. Social participatory sensing is closely related to social networking [76]. On the other hand, within opportunistic social sensing, the underlying system is in charge of acquiring needed data through relevant probes, as opposed to having the end-user providing them explicitly [38], [63], [36]. In urban sensing, also known as public sensing, data can be generated by everyone (or their devices) and exploited by everyone for public knowledge, including environment monitoring, or traffic updates [67]. In participatory urban sensing, users participate in providing information about the environment by exploiting the sensors/actuators embedded in their devices (which can be smartphones, vehicles, tablets, etc.) [67]. However data is only generated according to the owner's willingness to participate. Participatory urban sensing is especially characterized by scale issues at the data level, where data is generated by numerous individuals and should be processed and aggregated for knowledge to be inferred, involving adequate data scaling approaches [56]. Ikarus [96] is an example of participatory sensing, where data is collected by a large number of paragliders throughout their flights. The focus is on aggregating the data and rendering the results on a thermal map.

As outlined above, mobile social crowd-sensing has been a very active field of research for the last few years with various applications being targeted. However, effectively enabling mobile social crowd-sensing still raises a number of challenges, for which some early work may be identified:

- How to ensure that the system delivers the right quality of service, e.g., in terms of user-perceived delay, in spite of the resource constraints of mobile systems (e.g., [83])?
- How to guarantee the right level of privacy (e.g., [46], [85])?
- How to ensure the right level of participation from end-users so that mobile sensing indeed becomes a relevant source of accurate knowledge, which relates to eliciting adequate incentive mechanisms [98], in particular based on the understanding of mobile application usage [90], [89]?
- How to enrich sensor-generated content that is quantitative with user-generated one, thereby raising the issue of leveraging highly unstructured data while benefiting from a rich source of knowledge (e.g., sensing the crowdedness of a place combined with the feeling of people about the crowdedness, which may hint on the place's popularity as much as on discomfort)?

MOAIS Project-Team

3. Research Program

3.1. Scheduling

Participants: Pierre-François Dutot, Guillaume Huard, Grégory Mounié, Jean-Louis Roch, Denis Trystram, Frédéric Wagner.

The goal of this theme is to determine adequate multi-criteria objectives which are efficient (precision, reactivity, speed) and to study scheduling algorithms to reach these objectives.

In the context of parallel and distributed processing, the term *scheduling* is used with many acceptations. In general, scheduling means assigning tasks of a program (or processes) to the various components of a system (processors, communication links).

Researchers within MOAIS have been working on this subject for many years. They are known for their multiple contributions for determining the target dates and processors the tasks of a parallel program should be executed; especially regarding execution models (taking into account inter-task communications or any other system features) and the design of efficient algorithms (for which there exists a performance guarantee relative to the optimal scheduling).

Parallel tasks model and extensions. We have contributed to the definition and promotion of modern task models: parallel moldable tasks and divisible load. For both models, we have developed new techniques to derive efficient scheduling algorithms (with a good performance guaranty). We proposed recently some extensions taking into account machine unavailabilities (reservations).

Multi-objective Optimization. A natural question while designing practical scheduling algorithms is "which criterion should be optimized ?". Most existing works have been developed for minimizing the *makespan* (time of the latest tasks to be executed). This objective corresponds to a system administrator view who wants to be able to complete all the waiting jobs as soon as possible. The user, from his-her point of view, would be more interested in minimizing the average of the completion times (called *minsum*) of the whole set of submitted jobs. There exist several other objectives which may be pertinent for specific use. We worked on the problem of designing scheduling algorithms that optimize simultaneously several objectives with a theoretical guarantee on each objective. The main issue is that most of the policies are good for one criterion but bad for another one.
We have proposed an algorithm that is guaranteed for both *makespan* and *minsum*. This algorithm has been implemented for managing the resources of a cluster of the regional grid CIMENT. More recently, we extended such analysis to other objectives (makespan and reliability). We concentrate now on finding good algorithms able to schedule a set of jobs with a large variety of objectives simultaneously. For hard problems, we propose approximation of Pareto curves (best compromizes).

Incertainties. Most of the new execution supports are characterized by a higher complexity in predicting the parameters (high versatility in desktop grids, machine crash, communication congestion, cache effects, etc.). We studied some time ago the impact of incertainties on the scheduling algorithms. There are several ways for dealing with this problem: First, it is possible to design robust algorithms that can optimized a problem over a set of scenarii, another solution is to design flexible algorithms. Finally, we promote semi on-line approaches that start from an optimized off-line solution computed on an initial data set and updated during the execution on the "perturbed" data (stability analysis).

Game Theory. Game Theory is a framework that can be used for obtaining good solution of both previous problems (multi-objective optimization and incertain data). On the first hand, it can be used as a complement of multi-objective analysis. On the other hand, it can take into account the incertainties. We are curently working at formalizing the concept of cooperation.

Scheduling for optimizing parallel time and memory space. It is well known that parallel time and memory space are two antagonists criteria. However, for many scientific computations, the use of parallel architectures is motivated by increasing both the computation power and the memory space. Also, scheduling for optimizing both parallel time and memory space targets an important multicriteria objective. Based on the analysis of the dataflow related to the execution, we have proposed a scheduling algorithm with provable performance.

Coarse-grain scheduling of fine grain multithreaded computations on heterogeneous platforms. Designing multi-objective scheduling algorithms is a transversal problem. Work-stealing scheduling is well studied for fine grain multithreaded computations with a small critical time: the speed-up is asymptotically optimal. However, since the number of tasks to manage is huge, the control of the scheduling is expensive. We proposed a generalized lock-free cactus stack execution mechanism, to extend previous results, mainly from Cilk, based on the *work-first principle* for strict multi-threaded computations on SMPs to general multithreaded computations with dataflow dependencies. The main result is that optimizing the sequential local executions of tasks enables to amortize the overhead of scheduling. This distributed work-stealing scheduling algorithm has been implemented in **XKaapi**.

3.2. Adaptive Parallel and Distributed Algorithms Design

Participants: François Broquedis, Pierre-François Dutot, Thierry Gautier, Guillaume Huard, Bruno Raffin, Jean-Louis Roch, Denis Trystram, Frédéric Wagner.

This theme deals with the analysis and the design of algorithmic schemes that control (statically or dynamically) the grain of interactive applications.

The classical approach consists in setting in advance the number of processors for an application, the execution being limited to the use of these processors. This approach is restricted to a constant number of identical resources and for regular computations. To deal with irregularity (data and/or computations on the one hand; heterogeneous and/or dynamical resources on the other hand), an alternate approach consists in adapting the potential parallelism degree to the one suited to the resources. Two cases are distinguished:

- in the classical bottom-up approach, the application provides fine grain tasks; then those tasks are clustered to obtain a minimal parallel degree.
- the top-down approach (Cilk, Cilk+, TBB, Hood, Athapascan) is based on a work-stealing scheduling driven by idle resources. A local sequential depth-first execution of tasks is favored when recursive parallelism is available.

Ideally, a good parallel execution can be viewed as a flow of computations flowing through resources with no control overhead. To minimize control overhead, the application has to be adapted: a parallel algorithm on p resources is not efficient on $q < p$ resources. On one processor, the scheduler should execute a sequential algorithm instead of emulating a parallel one. Then, the scheduler should adapt to resource availability by changing its underlying algorithm. This first way of adapting granularity is implemented by XKaapi (default work-stealing schedule based on work-first principle).

However, this adaptation is restrictive. More generally, the algorithm should adapt itself at runtime to improve its performance by decreasing the overheads induced by parallelism, namely the arithmetic operations and communications. This motivates the development of new parallel algorithmic schemes that enable the scheduler to control the distribution between computation and communication (grain) in the application to find the good balance between parallelism and synchronizations. MOAIS has exhibited several techniques to manage adaptivity from an algorithmic point of view:

- amortization of the number of global synchronizations required in an iteration (for the evaluation of a stopping criterion);
- adaptive deployment of an application based on on-line discovery and performance measurements of communication links;
- generic recursive cascading of two kind of algorithms: a sequential one, to provide efficient executions on the local resource, and a parallel one that enables an idle resource to extract parallelism to dynamically suit the degree of parallelism to the available resources.

The generic underlying approach consists in finding a good mix of various algorithms, what is often called a "poly-algorithm". Particular instances of this approach are Atlas library (performance benchmark are used to decide at compile time the best block size and instruction interleaving for sequential matrix product) and FFTW library (at run time, the best recursive splitting of the FFT butterfly scheme is precomputed by dynamic programming). Both cases rely on pre-benchmarking of the algorithms. Our approach is more general in the sense that it also enables to tune the granularity at any time during execution. The objective is to develop processor oblivious algorithms: similarly to cache oblivious algorithms, we define a parallel algorithm as *processor-oblivious* if no program variable that depends on architecture parameters, such as the number or processors or their respective speeds, needs to be tuned to minimize the algorithm runtime.

We have applied this technique to develop processor oblivious algorithms for several applications with provable performance: iterated and prefix sum (partial sums) computations, stream computations (cipher and hd-video transformation), 3D image reconstruction (based on the concurrent usage of multi-core and GPU), loop computations with early termination.

Extensions concern the development of algorithms that are both cache and processor oblivious on heterogeneous processors. The processor algorithms proposed for prefix sums and segmentation of an array are cache oblivious too.

3.3. Interactivity

Participants: Vincent Danjean, Pierre-François Dutot, Thierry Gautier, Bruno Raffin, Jean-Louis Roch.

The goal of this theme is to develop approaches to tackle interactivity in the context of large scale distributed applications.

We distinguish two types of interactions. A user can interact with an application having only little insight about the internal details of the program running. This is typically the case for a virtual reality application where the user just manipulates 3D objects. We have a "user-in-the-loop". In opposite, we have an "expert -in-the-loop" if the user is an expert that knows the limits of the progam that is being executed and that he can interacts with it to steer the execution. This is the case for instance when the user can change some parameters during the execution to improve the convergence of a computation.

3.3.1. User-in-the-loop

Some applications, like virtual reality applications, must comply with interactivity constraints. The user should be able to observe and interact with the application with an acceptable reaction delay. To reach this goal the user is often ready to accept a lower level of details. To execute such application on a distributed architecture requires to balance the workload and activation frequency of the different tasks. The goal is to optimize CPU and network resource use to get as close as possible to the reactivity/level of detail the user expect.

Virtual reality environments significantly improve the quality of the interaction by providing advanced interfaces. The display surface provided by multiple projectors in CAVE -like systems for instance, allows a high resolution rendering on a large surface. Stereoscopic visualization gives an information of depth. Sound and haptic systems (force feedback) can provide extra information in addition to visualized data. However driving such an environment requires an important computation power and raises difficult issues of synchronization to maintain the overall application coherent while guaranteeing a good latency, bandwidth (or refresh rate) and level of details. We define the coherency as the fact that the information provided to the different user senses at a given moment are related to the same simulated time.

Today's availability of high performance commodity components including networks, CPUs as well as graphics or sound cards make it possible to build large clusters or grid environments providing the necessary resources to enlarge the class of applications that can aspire to an interactive execution. However the approaches usually used for mid size parallel machines are not adapted. Typically, there exist two different approaches to handle data exchange between the processes (or threads). The synchronous (or FIFO) approach ensures all messages sent are received in the order they were sent. In this case, a process cannot compute a new state if all incoming buffers do not store at least one message each. As a consequence, the application

refresh rate is driven by the slowest process. This can be improved if the user knows the relative speed of each module and specify a read frequency on each of the incoming buffers. This approach ensures a strong coherency but impact on latency. This is the approach commonly used to ensure the global coherency of the images displayed in multi-projector environments.The other approach, the asynchronous one, comes from sampling systems. The producer updates data in a shared buffer asynchronously read by the consumer. Some updates may be lost if the consumer is slower than the producer. The process refresh rates are therefore totally independent. Latency is improved as produced data are consumed as soon as possible, but no coherency is ensured. This approach is commonly used when coupling haptic and visualization systems. A fine tuning of the application usually leads to satisfactory results where the user does not experience major incoherences. However, in both cases, increasing the number of computing nodes quickly makes infeasible hand tuning to keep coherency and good performance.

We propose to develop techniques to manage a distributed interactive application regarding the following criteria :

- latency (the application reactivity);
- refresh rate (the application continuity);
- coherency (between the different components);
- level of detail (the precision of computations).

We developed a programming environment, called FlowVR, that enables the expression and realization of loosen but controlled coherency policies between data flows. The goal is to give users the possibility to express a large variety of coherency policies from a strong coherency based on a synchronous approach to an uncontrolled coherency based on an asynchronous approach. It enables the user to loosen coherency where it is acceptable, to improve asynchronism and thus performance. This approach maximizes the refresh rate and minimizes the latency given the coherency policy and a fixed level of details. It still requires the user to tune many parameters. In a second step, we are planning to explore auto-adaptive techniques that enable to decrease the number of parameters that must be user tuned. The goal is to take into account (possibly dynamically) user specified high level parameters like target latencies, bandwidths and levels of details, and to have the system automatically adapt to reach a trade-off given the user wishes and the resources available. Issues include multi-criterion optimizations, adaptive algorithmic schemes, distributed decision making, global stability and balance of the regulation effort.

3.3.2. Expert-in-the-loop

Some applications can be interactively guided by an expert who may give advices or answer specific questions to hasten a problem resolution. A theoretical framework has been developed in the last decade to define precisely the complexity of a problem when interactions with an expert is allowed. We are studying these interactive proof systems and interactive complexity classes in order to define efficient interactive algorithms dedicated to scheduling problems. This, in particular, applies to load-balancing of interactive simulations when a user interaction can generate a sudden surge of imbalance which could be easily predicted by an operator.

3.4. Adaptive middleware for code coupling and data movements

Participants: François Broquedis, Vincent Danjean, Thierry Gautier, Clément Pernet, Bruno Raffin, Jean-Louis Roch, Frédéric Wagner.

This theme deals with the design and implementation of programming interfaces in order to achieve an efficient coupling of distributed components.

The implementation of interactive simulation application requires to assemble together various software components and to ensure a semantic on the displayed result. To take into account functional aspects of the computation (inputs, outputs) as well as non functional aspects (bandwidth, latency, persistence), elementary actions (method invocation, communication) have to be coordinated in order to meet some performance objective (precision, quality, fluidity, *etc*). In such a context the scheduling algorithm plays an important role to adapt the computational power of a cluster architecture to the dynamic behavior due to the interactivity.

Whatever the scheduling algorithm is, it is fundamental to enable the control of the simulation. The purpose of this research theme is to specify the semantics of the operators that perform components assembling and to develop a prototype to experiment our proposals on real architectures and applications.

3.4.1. Application Programming Interface

The specification of an API to compose interactive simulation application requires to characterize the components and the interaction between components.The respect of causality between elementary events ensures, at the application level, that a reader will see the *last* write with respect to an order. Such a consistency should be defined at the level of the application to control the events ordered by a chain of causality. For instance, one of the result of Athapascan was to prove that a data flow consistency is more efficient than other ones because it generates fewer messages. Beyond causality based interactions, new models of interaction should be studied to capture non predictable events (delay of communication, capture of image) while ensuring a semantic.

Our methodology is based on the characterization of interactions required between components in the context of an interactive simulation application. For instance, criteria could be coherency of visualization, degree of interactivity. Beyond such characterization we hope to provide an operational semantic of interactions (at least well suited and understood by usage) and a cost model. Moreover they should be preserved by composition to predict the cost of an execution for part of the application.

The main result relies on a computable representation of the future of an execution; representations such as macro data flow are well suited because they explicit which data are required by a task. Such a representation can be built at runtime by an interpretation technique: the execution of a function call is differed by computing beforehand at runtime a graph of tasks that represents the (future) calls to execute.

3.4.2. Kernel for Asynchronous, Adaptive, Parallel and Interactive Application

Managing the complexity related to fine grain components and reaching high efficiency on a cluster architecture require to consider a dynamic behavior. Also, the runtime kernel is based on a representation of the execution: data flow graph with attributes for each node and efficient operators will be the basis for our software. This kernel has to be specialized for the considered applications. The low layer of the kernel has features to transfer data and to perform remote signalization efficiently. Well known techniques and legacy code have to be reused. For instance, multithreading, asynchronous invocation, overlapping of latency by computing, parallel communication and parallel algorithms for collective operations are fundamental techniques to reach performance. Because the choice of the scheduling algorithm depends on the application and the architecture, the kernel will provide an *causally connected representation* of the system that is running. This allows to specialize the computation of a good schedule of the data flow graph by providing algorithms (scheduling algorithms for instance) that compute on this (causally connected) representation: any modification of the representation is turned into a modification on the system (the parallel program under execution). Moreover, the kernel provides a set of basic operators to manipulate the graph (*e.g.* computes a partition from a schedule, remapping tasks, ...) to allow to control a distributed execution.

MUSE Team

3. Research Program

3.1. Active probing methods

We are developing methods that actively introduce probes in the network to discover properties of the connected devices and network segments. We are focusing in particular on methods to discover properties of home networks (connected devices and their types) and to distinguish if performance bottlenecks lie within the home network versus outside. Our goal is to develop adaptative methods that can leverage the collaboration of the set of available devices (including end-user devices and the home router, depending on which devices are running the measurement software).

3.2. Passive monitoring methods

This part our research develops methods that simply observe network traffic to infer the performance of networked applications and the location of performance bottlenecks, as well as to extract patterns of web content consumption. We are working on techniques to collect network traffic both at user's end-devices and at home routers. We also have access to network traffic traces collected on a campus network and on a large European broadband access provider.

3.3. Inferring user online experience

We are developing hybrid measurement methods that combine passive network measurement techniques to infer application performance with techniques from HCI to measure user perception. We will later use the resulting datasets to build models of user perception of network performance based only on data that we can obtain automatically from the user device or from user's traffic observed in the network.

3.4. Content summarisation

We are working on methods to summarise a set of reviews (for example, movie reviews from Rotten Tomatoes or IMDB; or restaurant reviews from Yelp) with a set of representative tags. Each tag is a sequence of two or three words. In parallel, we are building a mobile app that allows users to directly enter tags instead of free-text reviews.

MYRIADS Project-Team

3. Research Program

3.1. Introduction

The research activity within the MYRIADS team encompasses several areas: distributed systems, middleware and programming models. We have chosen to provide a brief presentation of some of the scientific foundations associated with them: autonomic computing, future internet and SOA, distributed operating systems, and unconventional/nature-inspired programming.

3.2. Autonomic Computing

During the past years the development of raw computing power coupled with the proliferation of computer devices has grown at exponential rates. This phenomenal growth along with the advent of the Internet have led to a new age of accessibility — to other people, other applications and others systems. It is not just a matter of numbers. This boom has also led to unprecedented levels of complexity for the design and the implementation of these applications and systems, and of the way they work together. The increasing system scale is reaching a level beyond human ability to master its complexity.

This points towards an inevitable need to automate many of the functions associated with computing today. Indeed we want to interact with applications and systems intuitively, and we want to be far less involved in running them. Ideally, we would like computing systems to entirely manage themselves.

IBM [58] has named its vision for the future of computing "autonomic computing." According to IBM this new computer paradigm means the design and implementation of computer systems, software, storage and support that must exhibit the following basic fundamentals:

- **Flexibility.** An autonomic computing system must configure and reconfigure itself under varying, even unpredictable, conditions.
- **Accessibility.** The nature of the autonomic system is that it is always on.
- **Transparency.** The system will perform its tasks and adapt to a user's needs without dragging the user into the intricacies of its workings.

In the Myriads team we will act to satisfy these fundamentals.

3.3. Future Internet and SOA

Traditional information systems were built by integrating applications into a communication framework, such as CORBA or with an Enterprise Application Integration system (EAI). Today, companies need to be able to reconfigure themselves; they need to be able to include other companies' business, split or externalize some of their works very quickly. In order to do this, the information systems should react and adapt very efficiently. EAIs approaches did not provide the necessary agility because they were too tightly coupled and a large part of business processes were "hard wired" into company applications.

Web services and Service Oriented Architectures (SOA) partly provide agility because in SOA business processes are completely separated from applications which can only be viewed as providing services through an interface. With SOA technologies it is easily possible to modify business processes, change, add or remove services.

However, SOA and Web services technologies are mainly market-driven and sometimes far from the state-of-the-art of distributed systems. Achieving dependability or being able to guarantee Service Level Agreement (SLA) needs much more agility of software elements. Dynamic adaptability features are necessary at many different levels (business processes, service composition, service discovery and execution) and should be coordinated. When addressing very large scale systems, autonomic behaviour of services and other parts of service oriented architectures is necessary.

SOAs will be part of the "Future Internet". The "Future Internet" will encompass traditional Web servers and browsers to support company and people interactions (Internet of services), media interactions, search systems, etc. It will include many appliances (Internet of things). The key research domains in this area are network research, cloud computing, Internet of services and advanced software engineering.

The Myriads team will address adaptability and autonomy of SOAs in the context of Grids, Clouds and at large scale.

3.4. Distributed Operating Systems

An operating system provides abstractions such as files, processes, sockets to applications so that programmers can design their applications independently of the computer hardware. At execution time, the operating system is in charge of finding and managing the hardware resources necessary to implement these abstractions in a secure way. It also manages hardware and abstract resource sharing between different users and programs.

A distributed operating system makes a network of computers appear as a single machine. The structure of the network and the heterogeneity of the computation nodes are hidden to users. Members of the Myriads team members have a long experience in the design and implementation of distributed operating systems, for instance in Kerrighed, Vigne, and XtreemOS projects.

The cloud computing model [43], [40] introduces new challenges in the organization of the information infrastructure: security, identity management, adaptation to the environment (costs). The organization of large IT infrastructures is also impacted as their internal data-centers, sometimes called private clouds, need to cooperate with resources and services provisioned from the cloud in order to cope with workload variations. The advent of cloud and green computing introduces new challenges in the domain of distributed operating systems: resources can be provisioned and released dynamically, the distribution of the computations on the resources must be reevaluated periodically in order to reduce power consumption and resource usage costs. Distributed cloud operating system must adapt to these new challenges in order to reduce cost and energy, for instance, through the redistribution of the applications and services on a smaller set of resources.

The Myriads team works on the design and implementation of system services at IaaS and PaaS levels to autonomously manage cloud and cloud federations resources and support collaboration between cloud users.

3.5. Unconventional/Nature-inspired Programming

Levering the computing services available on the Internet requires to revisit programming models, with the idea of expressing decentralised and autonomous behaviours (in particular self-repairing, self-adaptation). More concretely, composing services within large scale platforms calls for mechanisms to adequately discover and select services at run time, upon failure, or unexpected results.

Nature metaphors have been shown to provide adequate abstractions to build autonomic systems. Firstly, we want to explore nature metaphors, such as the chemical programming model as alternative programming models for expressing the interactions and coordination of services at large scale to build applications dynamically.

Within the *chemical* paradigm, a program is seen as a solution in which molecules (data) float and react together to produce new data according to rules (programs). Such a paradigm, implicitly parallel and distributed, appears to be a good candidate to express high level interactions of software components. The language naturally focus on the coordination of distributed autonomous entities. Thus, our first objective is to extend the semantics of chemical programs, in order to model not only a distributed execution of a service coordination, but also, the interactions between the different *molecules* within the Internet of Services (users, companies, services, advertisements, requests, ...). At present, a distributed implementation of the chemical paradigm does not exist. Our second objective is to develop the concepts and techniques required for such an implementation. While the paradigm exhibit several limitations regarding its run-time complexity, revisiting the model and studying its implementation over distributed platforms, and then showing its relevance in concrete settings (such as service coordination) may constitute an innovative research area.

PHOENIX Project-Team

3. Research Program

3.1. Design-Driven Software Development

Raising the level of abstraction beyond programming is a very active research topic involving a range of areas, including software engineering, programming languages and formal verification. The challenge is to allow design dimensions of a software system, both functional and non-functional, to be expressed in a high-level way, instead of being encoded with a programming language. Such design dimensions can then be leveraged to verify conformance properties and to generate programming support.

Our research on this topic is to take up this challenge with an approach inspired by programming languages, introducing a full-fledged language for designing software systems and processing design descriptions both for verification and code generation purposes. Our approach is also DSL-inspired in that it defines a conceptual framework to guide software development. Lastly, to make our approach practical to software developers, we introduce a methodology and a suite of tools covering the development life-cycle.

To raise the level of abstraction beyond programming, the key approaches are model-driven engineering and architecture description languages. A number of *architecture description languages* have been proposed; they are either (1) coupled with a programming language (*e.g.,* [36]), providing some level of abstraction above programming, or (2) integrated into a programming language (*e.g.,* [29], [37]), mixing levels of abstraction. Furthermore, these approaches poorly leverage architecture descriptions to support programming, they are crudely integrated into existing development environments, or they are solely used for verification purposes. *Model-driven software development* is another actively researched area. This approach often lacks code generation and verification support. Finally, most (if not all) approaches related to our research goal are *general purpose*; their universal nature provides little, if any, guidance to design a software system. This situation is a major impediment to both reasoning about a design artifact and generating programming support.

3.2. Integrating Non-Functional Concerns into Software Design

Most existing design approaches do not address non-functional concerns. When they do, they do not provide an approach to non-functional concerns that covers the entire development life-cycle. Furthermore, they usually are general purpose, impeding the use of non-functional declarations for verification and code generation. For example, the Architecture Analysis & Design Language (AADL) is a standard dedicated to real-time embedded systems [32]. AADL provides language constructs for the specification of software systems (*e.g.,* component, port) and their deployment on execution platforms (*e.g.,* thread, process, memory). Using AADL, designers specify non-functional aspects by adding properties on language constructs (*e.g.,* the period of a thread) or using language extensions such as the Error Model Annex. [0] The software design concepts of AADL are still rather general purpose and give little guidance to the designer.

Beyond offering a conceptual framework, our language-based approach provides an ideal setting to address non-functional properties (*e.g.,* performance, reliability, security, ...). Specifically, a design language can be enriched with non-functional declarations to pursue two goals: (1) expanding further the type of conformance that can be checked between the design of a software system and its implementation, and (2) enabling additional programming support and guidance.

We are investigating this idea by extending our design language with non-functional declarations. For example, we have addressed error handling [10], access conflicts to resources [34], and quality of service constraints [33].

[0] The Error Model Annex is a standardized AADL extension for the description of errors [38].

Following our approach to paradigm-oriented software development, non-functional declarations are verified at design time, they generate support that guides and constrains programming, they produce a runtime system that preserves invariants.

3.3. Human-driven Software Design

Knowledge of the human characteristics (individual, social and organizational) allow the design of complex system and artifacts for increasing their efficacy. In our approach of assistive computing, a main challenge is the integration of facets of Human Factors in order to design technology support adapted to user needs in term of ergonomic properties (acceptability, usability, utility etc) and delivered functionalities (oriented task under user abilities contraints).

We adapt this approach to improve the independent living and self-determination of users with cognitive impairments by developing a variety of orchestration scenarios of networked objects (hardware/software) to provide a pervasive support to their activities. Human factors methodologies are adopted in our approach with the direct purpose the reliability and efficiency of the performance of digital support systems in respect of objectives of health and well-being of the person (monitoring, evaluation, and rehabilitation).

Precisely, our methodologies are based on a closed iterative loop, as described in the figure below :

- Identifying the person needs in a natural situation (*i.e.,* desired but problematic activities) according to Human Factors Models of activity (*i.e.,* environmental constraints; social support networks - caregivers and family; person's abilities)
- Designing environmental support that will assist the users to bypass their cognitive impairment (according to environmental models of cognitive compensatory mechanisms); and then implement this support in terms of technological solutions (scenarios of networked objects, hardware interface, software interface, interaction style, *etc*)
- Empirically evaluating the assistive solution based on human experimentations that includes ergonomic assessments (acceptability, usability, usefulness, *etc*) as well as longitudinal evaluations of use's efficacy in terms of activities performed by the individual, of satisfaction and well-being provided to the individual but also to his/her entourage (family and caregivers).

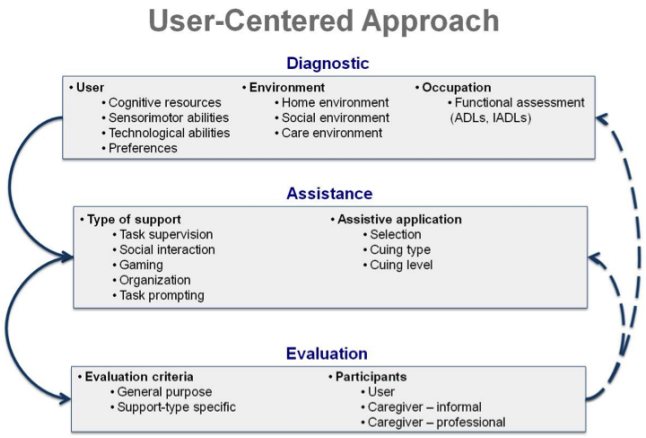


Figure 1. User-Centred Approach

RAP Project-Team

3. Research Program

3.1. Design and Analysis of Algorithms

Data Structures, Stochastic Algorithms

The general goal of the research in this domain is of designing algorithms to analyze and control the traffic of communication networks. The team is currently involved in the design of algorithms to allocate bandwidth in optical networks and also to allocate resources in large distributed networks. See the corresponding sections below.

The team also pursues analysis of algorithms and data structures in the spirit of the former Algorithms team. The team is especially interested in the ubiquitous divide-and-conquer paradigm and its applications to the design of search trees, and stable collision resolution protocols.

3.2. Scaling of Markov Processes

The growing complexity of communication networks makes it more difficult to apply classical mathematical methods. For a one/two-dimensional Markov process describing the evolution of some network, it is sometimes possible to write down the equilibrium equations and to solve them. The key idea to overcome these difficulties is to consider the system in limit regimes. This list of possible renormalization procedures is, of course, not exhaustive. The advantages of these methods lie in their flexibility to various situations and to the interesting theoretical problems they raised.

A fluid limit scaling is a particularly important means to scale a Markov process. It is related to the first order behavior of the process and, roughly speaking, amounts to a functional law of large numbers for the system considered.

A fluid limit keeps the main characteristics of the initial stochastic process while some second order stochastic fluctuations disappear. In “good” cases, a fluid limit is a deterministic function, obtained as the solution of some ordinary differential equation. As can be expected, the general situation is somewhat more complicated. These ideas of rescaling stochastic processes have emerged recently in the analysis of stochastic networks, to study their ergodicity properties in particular.

3.3. Structure of random networks

This line of research aims at understanding the global structure of stochastic networks (connectivity, magnitude of distances, etc) via models of random graphs. It consists of two complementary foundational and applied aspects of connectivity.

RANDOM GRAPHS, STATISTICAL PHYSICS AND COMBINATORIAL OPTIMIZATION. The connectivity of usual models for networks based on random graphs models (Erdős–Rényi and random geometric graphs) may be tuned by adjusting the average degree. There is a *phase transition* as the average degree approaches one, a *giant* connected component containing a positive proportion of the nodes suddenly appears. The phase of practical interest is the *supercritical* one, when there is at least a giant component, while the theoretical interest lies at the *critical phase*, the break-point just before it appears.

At the critical point there is not yet a macroscopic component and the network consists of a large number of connected component at the mesoscopic scale. From a theoretical point of view, this phase is most interesting since the structure of the clusters there is expected (heuristically) to be *universal*. Understanding this phase and its universality is a great challenge that would impact the knowledge of phase transitions in all high-dimensional models of *statistical physics* and *combinatorial optimization*.

RANDOM GEOMETRIC GRAPHS AND WIRELESS NETWORKS. The level of connection of the network is of course crucial, but the *scalability* imposes that the underlying graph also be *sparse*: trade offs must be made, which required a fine evaluation of the costs/benefits. Various direct and indirect measures of connectivity are crucial to these choices: What is the size of the overwhelming connected component? When does complete connectivity occur? What is the order of magnitude of distances? Are paths to a target easy to find using only local information? Are there simple broadcasting algorithms? Can one put an end to viral infections? How much time for a random crawler to see most of the network?

NAVIGATION AND POINT LOCATION IN RANDOM MESHES. Other applications which are less directly related to networks include the design of improved navigation or point location algorithms in geometric meshes such as the Delaunay triangulation build from random point sets. There the graph model is essentially fixed, but the constraints it imposes raise a number of challenging problems. The aim is to prove performance guarantees for these algorithms which are used in most manipulations of the meshes.

REGAL Project-Team

3. Research Program

3.1. Research rationale

As society relies more and more on computers, responsiveness, correctness and security are increasingly critical. At the same time, systems are growing larger, more parallel, and more unpredictable. Our research agenda is to design Computer Systems that remain correct and efficient despite this increased complexity and in spite of conflicting requirements. The term *"Computer Systems"* is interpreted broadly,[0] and includes system architecture, operating systems, distributed systems, multiprocessor systems, and touches on related areas such as computer networks, distributed databases or support for big data. The interests of the Regal group cover the whole spectrum from theory to experimentation, with a strong focus on algorithm design and implementation.

This holistic approach allows us to address related problems at different levels. It also permits us to efficiently share knowledge and expertise, and is a source of originality.

Computer Systems is a rapidly evolving domain, with strong interactions with industry. Two main evolutions in the Computer Systems area have strongly influenced our research activities:

3.1.1. Modern computer systems are increasingly parallel and distributed.

Ensuring the persistence, availability and consistency of data in a distributed setting is a major requirement: the system must remain correct despite slow networks, disconnection, crashes, failures, churn, and attacks. Ease of use, performance and efficiency are equally important for systems to be accepted. These requirements are somewhat conflicting, and there are many algorithmic and engineering trade-offs, which often depend on specific workloads or usage scenarios.

Years of research in distributed systems are now coming to fruition, and are being used by millions of users of web systems, peer-to-peer systems, gaming and social applications, or cloud computing. These new usages bring new challenges of extreme scalability and adaptation to dynamically-changing conditions, where knowledge of system state can only be partial and incomplete. The challenges of distributed computing listed above are subject to new trade-offs.

Innovative environments that motivate our research include cloud computing, geo-replication, edge clouds, peer-to-peer (P2P) systems, dynamic networks, and manycore machines. The scientific challenges are scalability, fault tolerance, security, dynamicity and the virtualization of the physical infrastructure. Algorithms designed for classical distributed systems, such as resource allocation, data storage and placement, and concurrent and consistent access to shared data, need to be revisited to work properly under the constraints of these new environments.

Regal focuses in particular on two key challenges in these areas: the adaptation of algorithms to the new dynamics of distributed systems and data management on large configurations.

3.1.2. Multicore architectures are everywhere.

The fine-grained parallelism offered by multicore architectures has the potential to open highly parallel computing to new application areas. To make this a reality, however, many issues, including issues that have previously arisen in distributed systems, need to be addressed. Challenges include obtaining a consistent view of shared resources, such as memory, and optimally distributing computations among heterogeneous architectures, such as CPUs, GPUs, and other specialized processors. As compared to distributed systems, in the case of multicore architectures, these issues arise at a more fine-grained level, leading to the need for different solutions and different cost-benefit trade-offs.

[0] This follows the definition from the journal of reference in our field, ACM Transactions on Computer Systems.

Of particular interest to Regal are topics related to memory management in high-end multicore computers, such as garbage collection of very large memories and system support for massive databases of highly-structured data.

RMOD Project-Team

3. Research Program

3.1. Software Reengineering

Strong coupling among the parts of an application severely hampers its evolution. Therefore, it is crucial to answer the following questions: How to support the substitution of certain parts while limiting the impact on others? How to identify reusable parts? How to modularize an object-oriented application?

Having good classes does not imply a good application layering, absence of cycles between packages and reuse of well-identified parts. Which notion of cohesion makes sense in presence of late-binding and programming frameworks? Indeed, frameworks define a context that can be extended by subclassing or composition: in this case, packages can have a low cohesion without being a problem for evolution. How to obtain algorithms that can be used on real cases? Which criteria should be selected for a given remodularization?

To help us answer these questions, we work on enriching Moose, our reengineering environment, with a new set of analyses [56], [55]. We decompose our approach in three main and potentially overlapping steps:

1. Tools for understanding applications,
2. Remodularization analyses,
3. Software Quality.

3.1.1. Tools for understanding applications

Context and Problems. We are studying the problems raised by the understanding of applications at a larger level of granularity such as packages or modules. We want to develop a set of conceptual tools to support this understanding.

Some approaches based on Formal Concept Analysis (FCA) [84] show that such an analysis can be used to identify modules. However the presented examples are too small and not representative of real code.

Research Agenda.

FCA provides an important approach in software reengineering for software understanding, design anomalies detection and correction, but it suffers from two problems: (i) it produces lattices that must be interpreted by the user according to his/her understanding of the technique and different elements of the graph; and, (ii) the lattice can rapidly become so big that one is overwhelmed by the mass of information and possibilities [45]. We look for solutions to help people putting FCA to real use.

3.1.2. Remodularization analyses

Context and Problems. It is a well-known practice to layer applications with bottom layers being more stable than top layers [72]. Until now, few works have attempted to identify layers in practice: Mudpie [86] is a first cut at identifying cycles between packages as well as package groups potentially representing layers. DSM (dependency structure matrix) [85], [80] seems to be adapted for such a task but there is no serious empirical experience that validates this claim. From the side of remodularization algorithms, many were defined for procedural languages [68]. However, object-oriented programming languages bring some specific problems linked with late-binding and the fact that a package does not have to be systematically cohesive since it can be an extension of another one [87], [59].

As we are designing and evaluating algorithms and analyses to remodularize applications, we also need a way to understand and assess the results we are obtaining.

Research Agenda. We work on the following items:

Layer identification. We propose an approach to identify layers based on a semi-automatic classification of package and class interrelationships that they contain. However, taking into account the wish or knowledge of the designer or maintainer should be supported.

Cohesion Metric Assessment. We are building a validation framework for cohesion/coupling metrics to determine whether they actually measure what they promise to. We are also compiling a number of traditional metrics for cohesion and coupling quality metrics to evaluate their relevance in a software quality setting.

3.1.3. Software Quality

Research Agenda. Since software quality is fuzzy by definition and a lot of parameters should be taken into account we consider that defining precisely a unique notion of software quality is definitively a Grail in the realm of software engineering. The question is still relevant and important. We work on the two following items:

Quality models. We studied existing quality models and the different options to combine indicators — often, software quality models happily combine metrics, but at the price of losing the explicit relationships between the indicator contributions. There is a need to combine the results of one metric over all the software components of a system, and there is also the need to combine different metric results for any software component. Different combination methods are possible that can give very different results. It is therefore important to understand the characteristics of each method.

Bug prevention. Another aspect of software quality is validating or monitoring the source code to avoid the emergence of well known sources of errors and bugs. We work on how to best identify such common errors, by trying to identify earlier markers of possible errors, or by helping identifying common errors that programmers did in the past.

3.2. Language Constructs for Modular Design

While the previous axis focuses on how to help remodularizing existing software, this second research axis aims at providing new language constructs to build more flexible and recomposable software. We will build on our work on traits [82], [57] and classboxes [46] but also start to work on new areas such as isolation in dynamic languages. We will work on the following points: (1) Traits and (2) Modularization as a support for isolation.

3.2.1. Traits-based program reuse

Context and Problems. Inheritance is well-known and accepted as a mechanism for reuse in object-oriented languages. Unfortunately, due to the coarse granularity of inheritance, it may be difficult to decompose an application into an optimal class hierarchy that maximizes software reuse. Existing schemes based on single inheritance, multiple inheritance, or mixins, all pose numerous problems for reuse.

To overcome these problems, we designed a new composition mechanism called Traits [82], [57]. Traits are pure units of behavior that can be composed to form classes or other traits. The trait composition mechanism is an alternative to multiple or mixin inheritance in which the composer has full control over the trait composition. The result enables more reuse than single inheritance without introducing the drawbacks of multiple or mixin inheritance. Several extensions of the model have been proposed [54], [76], [47], [58] and several type systems were defined [60], [83], [77], [70].

Traits are reusable building blocks that can be explicitly composed to share methods across unrelated class hierarchies. In their original form, traits do not contain state and cannot express visibility control for methods. Two extensions, stateful traits and freezable traits, have been proposed to overcome these limitations. However, these extensions are complex both to use for software developers and to implement for language designers.

Research Agenda: Towards a pure trait language. We plan distinct actions: (1) a large application of traits, (2) assessment of the existing trait models and (3) bootstrapping a pure trait language.

- To evaluate the expressiveness of traits, some hierarchies were refactored, showing code reuse [49]. However, such large refactorings, while valuable, may not exhibit all possible composition problems, since the hierarchies were previously expressed using single inheritance and following certain patterns. We want to redesign from scratch the collection library of Smalltalk (or part of it). Such a redesign should on the one hand demonstrate the added value of traits on a real large and redesigned library and on the other hand foster new ideas for the bootstrapping of a pure trait-based language.

  In particular we want to reconsider the different models proposed (stateless [57], stateful [48], and freezable [58]) and their operators. We will compare these models by (1) implementing a trait-based collection hierarchy, (2) analyzing several existing applications that exhibit the need for traits. Traits may be flattened [75]. This is a fundamental property that confers to traits their simplicity and expressiveness over Eiffel's multiple inheritance. Keeping these aspects is one of our priority in forthcoming enhancements of traits.
- Alternative trait models. This work revisits the problem of adding state and visibility control to traits. Rather than extending the original trait model with additional operations, we use a fundamentally different approach by allowing traits to be lexically nested within other modules. This enables traits to express (shared) state and visibility control by hiding variables or methods in their lexical scope. Although the traits' "flattening property" no longer holds when they can be lexically nested, the combination of traits with lexical nesting results in a simple and more expressive trait model. We formally specify the operational semantics of this combination. Lexically nested traits are fully implemented in AmbientTalk, where they are used among others in the development of a Morphic-like UI framework.
- We want to evaluate how inheritance can be replaced by traits to form a new object model. For this purpose we will design a minimal reflective kernel, inspired first from ObjVlisp [53] then from Smalltalk [63].

3.2.2. Reconciling Dynamic Languages and Isolation

Context and Problems. More and more applications require dynamic behavior such as modification of their own execution (often implemented using reflective features [67]). For example, F-script allows one to script Cocoa Mac-OS X applications and Lua is used in Adobe Photoshop. Now in addition more and more applications are updated on the fly, potentially loading untrusted or broken code, which may be problematic for the system if the application is not properly isolated. Bytecode checking and static code analysis are used to enable isolation, but such approaches do not really work in presence of dynamic languages and reflective features. Therefore there is a tension between the need for flexibility and isolation.

Research Agenda: Isolation in dynamic and reflective languages. To solve this tension, we will work on *Sure*, a language where isolation is provided by construction: as an example, if the language does not offer field access and its reflective facilities are controlled, then the possibility to access and modify private data is controlled. In this context, layering and modularizing the meta-level [50], as well as controlling the access to reflective features [51], [52] are important challenges. We plan to:

- Study the isolation abstractions available in erights (http://www.erights.org) [74], [73], and Java's class loader strategies [69], [64].
- Categorize the different reflective features of languages such as CLOS [66], Python and Smalltalk [78] and identify suitable isolation mechanisms and infrastructure [61].
- Assess different isolation models (access rights, capabilities [79]...) and identify the ones adapted to our context as well as different access and right propagation.
- Define a language based on
  - the decomposition and restructuring of the reflective features [50],

- the use of encapsulation policies as a basis to restrict the interfaces of the controlled objects [81],
- the definition of method modifiers to support controlling encapsulation in the context of dynamic languages.

An open question is whether, instead of providing restricted interfaces, we could use traits to grant additional behavior to specific instances: without trait application, the instances would only exhibit default public behavior, but with additional traits applied, the instances would get extra behavior. We will develop *Sure*, a modular extension of the reflective kernel of Smalltalk (since it is one of the languages offering the largest set of reflective features such as pointer swapping, class changing, class definition...) [78].

ROMA Team

3. Research Program

3.1. Algorithms for probabilistic environments

There are two main research directions under this research theme. In the first one, we consider the problem of the efficient execution of applications in a failure-prone environment. Here, probability distributions are used to describe the potential behavior of computing platforms, namely when hardware components are subject to faults. In the second research direction, probability distributions are used to describe the characteristics and behavior of applications.

3.1.1. Application resilience

An application is resilient if it can successfully produce a correct result in spite of potential faults in the underlying system. Application resilience can involve a broad range of techniques, including fault prediction, error detection, error containment, error correction, checkpointing, replication, migration, recovery, etc. Faults are quite frequent in the most powerful existing supercomputers. The Jaguar platform, which ranked third in the TOP 500 list in November 2011 [42], had an average of 2.33 faults per day during the period from August 2008 to February 2010 [66]. The mean-time between faults of a platform is inversely proportional to its number of components. Progresses will certainly be made in the coming years with respect to the reliability of individual components. However, designing and building high-reliability hardware components is far more expensive than using lower reliability top-of-the-shelf components. Furthermore, low-power components may not be available with high-reliability. Therefore, it is feared that the progresses in reliability will far from compensate the steady projected increase of the number of components in the largest supercomputers. Already, application failures have a huge computational cost. In 2008, the DARPA white paper on "System resilience at extreme scale" [41] stated that high-end systems wasted 20% of their computing capacity on application failure and recovery.

In such a context, any application using a significant fraction of a supercomputer and running for a significant amount of time will have to use some fault-tolerance solution. It would indeed be unacceptable for an application failure to destroy centuries of CPU-time (some of the simulations run on the Blue Waters platform consumed more than 2,700 years of core computing time [37] and lasted over 60 hours; the most time-consuming simulations of the US Department of Energy (DoE) run for weeks to months on the most powerful existing platforms [40]).

Our research on resilience follows two different directions. On the one hand we design new resilience solutions, either generic fault-tolerance solutions or algorithm-based solutions. On the other hand we model and theoretically analyze the performance of existing and future solutions, in order to tune their usage and help determine which solution to use in which context.

3.1.2. Scheduling strategies for applications with a probabilistic behavior

Static scheduling algorithms are algorithms where all decisions are taken before the start of the application execution. On the contrary, in non-static algorithms, decisions may depend on events that happen during the execution. Static scheduling algorithms are known to be superior to dynamic and system-oriented approaches in stable frameworks [47], [53], [54], [65], that is, when all characteristics of platforms and applications are perfectly known, known a priori, and do not evolve during the application execution. In practice, the prediction of application characteristics may be approximative or completely infeasible. For instance, the amount of computations and of communications required to solve a given problem in parallel may strongly depend on some input data that are hard to analyze (this is for instance the case when solving linear systems using full pivoting).

We plan to consider applications whose characteristics change dynamically and are subject to uncertainties. In order to benefit nonetheless from the power of static approaches, we plan to model application uncertainties and variations through probabilistic models, and to design for these applications scheduling strategies that are either static, or partially static and partially dynamic.

3.2. Platform-aware scheduling strategies

In this theme, we study and design scheduling strategies, focusing either on energy consumption or on memory behavior. In other words, when designing and evaluating these strategies, we do not limit our view to the most classical platform characteristics, that is, the computing speed of cores and accelerators, and the bandwidth of communication links.

In most existing studies, a single optimization objective is considered, and the target is some sort of absolute performance. For instance, most optimization problems aim at the minimization of the overall execution time of the application considered. Such an approach can lead to a very significant waste of resources, because it does not take into account any notion of efficiency nor of yield. For instance, it may not be meaningful to use twice as many resources just to decrease by 10% the execution time. In all our work, we plan to look only for algorithmic solutions that make a "clever" usage of resources. However, looking for the solution that optimizes a metric such as the efficiency, the energy consumption, or the memory-peak minimization, is doomed for the type of applications we consider. Indeed, in most cases, any optimal solution for such a metric is a sequential solution, and sequential solutions have prohibitive execution times. Therefore, it becomes mandatory to consider multi-criteria approaches where one looks for trade-offs between some user-oriented metrics that are typically related to notions of Quality of Service—execution time, response time, stretch, throughput, latency, reliability, etc.—and some system-oriented metrics that guarantee that resources are not wasted. In general, we will not look for the Pareto curve, that is, the set of all dominating solutions for the considered metrics. Instead, we will rather look for solutions that minimize some given objective while satisfying some bounds, or "budgets", on all the other objectives.

3.2.1. Energy-aware algorithms

Energy-aware scheduling has proven an important issue in the past decade, both for economical and environmental reasons. Energy issues are obvious for battery-powered systems. They are now also important for traditional computer systems. Indeed, the design specifications of any new computing platform now always include an upper bound on energy consumption. Furthermore, the energy bill of a supercomputer may represent a significant share of its cost over its lifespan.

Technically, a processor running at speed s dissipates s^α watts per unit of time with $2 \leq \alpha \leq 3$ [45], [46], [51]; hence, it consumes $s^\alpha \times d$ joules when operated during d units of time. Therefore, energy consumption can be reduced by using speed scaling techniques. However it was shown in [67] that reducing the speed of a processor increases the rate of transient faults in the system. The probability of faults increases exponentially, and this probability cannot be neglected in large-scale computing [61]. In order to make up for the loss in *reliability* due to the energy efficiency, different models have been proposed for fault tolerance: (i) *re-execution* consists in re-executing a task that does not meet the reliability constraint [67]; (ii) *replication* consists in executing the same task on several processors simultaneously, in order to meet the reliability constraints [44]; and (iii) *checkpointing* consists in "saving" the work done at some certain instants, hence reducing the amount of work lost when a failure occurs [60].

Energy issues must be taken into account at all levels, including the algorithm-design level. We plan to both evaluate the energy consumption of existing algorithms and to design new algorithms that minimize energy consumption using tools such as resource selection, dynamic frequency and voltage scaling, or powering-down of hardware components.

3.2.2. Memory-aware algorithms

For many years, the bandwidth between memories and processors has increased more slowly than the computing power of processors, and the latency of memory accesses has been improved at an even slower

pace. Therefore, in the time needed for a processor to perform a floating point operation, the amount of data transferred between the memory and the processor has been decreasing with each passing year. The risk is for an application to reach a point where the time needed to solve a problem is no longer dictated by the processor computing power but by the memory characteristics, comparable to the *memory wall* that limits CPU performance. In such a case, processors would be greatly under-utilized, and a large part of the computing power of the platform would be wasted. Moreover, with the advent of multicore processors, the amount of memory per core has started to stagnate, if not to decrease. This is especially harmful to memory intensive applications. The problems related to the sizes and the bandwidths of memories are further exacerbated on modern computing platforms because of their deep and highly heterogeneous hierarchies. Such a hierarchy can extend from core private caches to shared memory within a CPU, to disk storage and even tape-based storage systems, like in the Blue Waters supercomputer [38]. It may also be the case that heterogeneous cores are used (such as hybrid CPU and GPU computing), and that each of them has a limited memory.

Because of these trends, it is becoming more and more important to precisely take memory constraints into account when designing algorithms. One must not only take care of the amount of memory required to run an algorithm, but also of the way this memory is accessed. Indeed, in some cases, rather than to minimize the amount of memory required to solve the given problem, one will have to maximize data reuse and, especially, to minimize the amount of data transferred between the different levels of the memory hierarchy (minimization of the volume of memory inputs-outputs). This is, for instance, the case when a problem cannot be solved by just using the in-core memory and that any solution must be out-of-core, that is, must use disks as storage for temporary data.

It is worth noting that the cost of moving data has lead to the development of so called "communication-avoiding algorithms" [57]. Our approach is orthogonal to these efforts: in communication-avoiding algorithms, the application is modified, in particular some redundant work is done, in order to get rid of some communication operations, whereas in our approach, we do not modify the application, which is provided as a task graph, but we minimize the needed memory peak only by carefully scheduling tasks.

3.3. High-performance computing and linear algebra

Our work on high-performance computing and linear algebra is organized along three research directions. The first direction is devoted to direct solvers of sparse linear systems. The second direction is devoted to combinatorial scientific computing, that is, the design of combinatorial algorithms and tools that solve problems encountered in some of the other research themes, like the problems faced in the preprocessing phases of sparse direct solvers. The last direction deals with the adaptation of classical dense linear algebra kernels to the architecture of future computing platforms.

3.3.1. Direct solvers for sparse linear systems

The solution of sparse systems of linear equations (symmetric or unsymmetric, often with an irregular structure, from a few hundred thousand to a few hundred million equations) is at the heart of many scientific applications arising in domains such as geophysics, structural mechanics, chemistry, electromagnetism, numerical optimization, or computational fluid dynamics, to cite a few. The importance and diversity of applications are a main motivation to pursue research on sparse linear solvers. Because of this wide range of applications, any significant progress on solvers will have a significant impact in the world of simulation. Research on sparse direct solvers in general is very active for the following main reasons:

- many applications fields require large-scale simulations that are still too big or too complicated with respect to today's solution methods;
- the current evolution of architectures with massive, hierarchical, multicore parallelism imposes to overhaul all existing solutions, which represents a major challenge for algorithm and software development;
- the evolution of numerical needs and types of simulations increase the importance, frequency, and size of certain classes of matrices, which may benefit from a specialized processing (rather than resort to a generic one).

Our research in the field is strongly related to the software package MUMPS (see Section 5.1). MUMPS is both an experimental platform for academics in the field of sparse linear algebra, and a software package that is widely used in both academia and industry. The software package MUMPS enables us to (i) confront our research to the real world, (ii) develop contacts and collaborations, and (iii) receive continuous feedback from real-life applications, which is extremely critical to validate our research work. The feedback from a large user community also enables us to direct our long-term objectives towards meaningful directions.

In this context, we aim at designing parallel sparse direct methods that will scale to large modern platforms, and that are able to answer new challenges arising from applications, both efficiently—from a resource consumption point of view—and accurately—from a numerical point of view. For that, and even with increasing parallelism, we do not want to sacrifice in any manner numerical stability, based on threshold partial pivoting, one of the main originalities of our approach (our "trademark") in the context of direct solvers for distributed-memory computers; although this makes the parallelization more complicated, applying the same pivoting strategy as in the serial case ensures numerical robustness of our approach, which we generally measure in terms of sparse backward error. In order to solve the hard problems resulting from the always-increasing demands in simulations, special attention must also necessarily be paid to memory usage (and not only execution time). This requires specific algorithmic choices and scheduling techniques. From a complementary point of view, it is also necessary to be aware of the functionality requirements from the applications and from the users, so that robust solutions can be proposed for a wide range of applications.

Among direct methods, we rely on the multifrontal method [55], [56], [59]. This method usually exhibits a good data locality and hence is efficient in cache-based systems. The task graph associated with the multifrontal method is in the form of a tree whose characteristics should be exploited in a parallel implementation.

Our work is organized along two main research directions. In the first one we aim at efficiently addressing new architectures that include massive, hierarchical parallelism. In the second one, we aim at reducing the running time complexity and the memory requirements of direct solvers, while controlling accuracy.

3.3.2. Combinatorial scientific computing

Combinatorial scientific computing (CSC) is a recently coined term (circa 2002) for interdisciplinary research at the intersection of discrete mathematics, computer science, and scientific computing. In particular, it refers to the development, application, and analysis of combinatorial algorithms to enable scientific computing applications. CSC's deepest roots are in the realm of direct methods for solving sparse linear systems of equations where graph theoretical models have been central to the exploitation of sparsity, since the 1960s. The general approach is to identify performance issues in a scientific computing problem, such as memory use, parallel speed up, and/or the rate of convergence of a method, and to develop combinatorial algorithms and models to tackle those issues.

Our target scientific computing applications are (i) the preprocessing phases of direct methods (in particular MUMPS), iterative methods, and hybrid methods for solving linear systems of equations; and (ii) the mapping of tasks (mostly the sub-tasks of the mentioned solvers) onto modern computing platforms. We focus on the development and use of graph and hypergraph models, and related tools such as hypergraph partitioning algorithms, to solve problems of load balancing and task mapping. We also focus on bipartite graph matching and vertex ordering methods for reducing the memory overhead and computational requirements of solvers. Although we direct our attention on these models and algorithms through the lens of linear system solvers, our solutions are general enough to be applied to some other resource optimization problems.

3.3.3. Dense linear algebra on post-petascale multicore platforms

The quest for efficient, yet portable, implementations of dense linear algebra kernels (QR, LU, Cholesky) has never stopped, fueled in part by each new technological evolution. First, the LAPACK library [49] relied on BLAS level 3 kernels (Basic Linear Algebra Subroutines) that enable to fully harness the computing power of a single CPU. Then the SCALAPACK library [48] built upon LAPACK to provide a coarse-grain parallel version, where processors operate on large block-column panels. Inter-processor communications occur through highly tuned MPI send and receive primitives. The advent of multi-core processors has led to a

major modification in these algorithms [50], [64], [58]. Each processor runs several threads in parallel to keep all cores within that processor busy. Tiled versions of the algorithms have thus been designed: dividing large block-column panels into several tiles allows for a decrease in the granularity down to a level where many smaller-size tasks are spawned. In the current panel, the diagonal tile is used to eliminate all the lower tiles in the panel. Because the factorization of the whole panel is now broken into the elimination of several tiles, the update operations can also be partitioned at the tile level, which generates many tasks to feed all cores.

The number of cores per processor will keep increasing in the following years. It is projected that high-end processors will include at least a few hundreds of cores. This evolution will require to design new versions of libraries. Indeed, existing libraries rely on a static distribution of the work: before the beginning of the execution of a kernel, the location and time of the execution of all of its component is decided. In theory, static solutions enable to precisely optimize executions, by taking parameters like data locality into account. At run time, these solutions proceed at the pace of the slowest of the cores, and they thus require a perfect load-balancing. With a few hundreds, if not a thousand, cores per processor, some tiny differences between the computing times on the different cores ("jitter") are unavoidable and irremediably condemn purely static solutions. Moreover, the increase in the number of cores per processor once again mandates to increase the number of tasks that can be executed in parallel.

We study solutions that are part-static part-dynamic, because such solutions have been shown to outperform purely dynamic ones [52]. On the one hand, the distribution of work among the different nodes will still be statically defined. On the other hand, the mapping and the scheduling of tasks inside a processor will be dynamically defined. The main difficulty when building such a solution will be to design lightweight dynamic schedulers that are able to guarantee both an excellent load-balancing and a very efficient use of data locality.

RUNTIME Team

3. Research Program

3.1. Runtime Systems Evolution

parallel,distributed,cluster,environment,library,communication,multithreading,multicore

This research project takes place within the context of high-performance computing. It seeks to contribute to the design and implementation of parallel runtime systems that shall serve as a basis for the implementation of high-level parallel middleware. Today, the implementation of such software (programming environments, numerical libraries, parallel language compilers, parallel virtual machines, etc.) has become so complex that the use of portable, low-level runtime systems is unavoidable.

Our research project centers on three main directions:

Mastering large, hierarchical multiprocessor machines With the beginning of the new century, computer makers have initiated a long term move of integrating more and more processing units, as an answer to the frequency wall hit by the technology. This integration cannot be made in a basic, planar scheme beyond a couple of processing units for scalability reasons. Instead, vendors have to resort to organize those processing units following some hierarchical structure scheme. A level in the hierarchy is then materialized by small groups of units sharing some common local cache or memory bank. Memory accesses outside the locality of the group are still possible thanks to bus-level consistency mechanisms but are significantly more expensive than local accesses, which, by definition, characterizes NUMA architectures.

Thus, the task scheduler must feed an increasing number of processing units with work to execute and data to process while keeping the rate of penalized memory accesses as low as possible. False sharing, ping-pong effects, data vs task locality mismatches, and even task vs task locality mismatches between tightly synchronizing activities are examples of the numerous sources of overhead that may arise if threads and data are not distributed properly by the scheduler. To avoid these pitfalls, the scheduler therefore needs accurate information both about the computing platform layout it is running on and about the structure and activities relationships of the application it is scheduling.

As quoted by Gao *et al.* [46], we believe it is important to expose domain-specific knowledge semantics to the various software components in order to organize computation according to the application and architecture. Indeed, the whole software stack, from the application to the scheduler, should be involved in the parallelizing, scheduling and locality adaptation decisions by providing useful information to the other components. Unfortunately, most operating systems only provide a poor scheduling API that does not allow applications to transmit valuable *hints* to the system.

This is why we investigate new approaches in the design of thread schedulers, focusing on high-level abstractions to both model hierarchical architectures and describe the structure of applications' parallelism. In particular, we have introduced the *bubble* scheduling concept [7] that helps to structure relations between threads in a way that can be efficiently exploited by the underlying thread scheduler. *Bubbles* express the inherent parallel structure of multithreaded applications: they are abstractions for grouping threads which "work together" in a recursive way. We are exploring how to dynamically schedule these irregular nested sets of threads on hierarchical machines [3], the key challenge being to schedule related threads as closely as possible in order to benefit from cache effects and avoid NUMA penalties. We are also exploring how to improve the transfer of scheduling hints from the programming environment to the runtime system, to achieve better computation efficiency.

This is also the reason why we explore new languages and compiler optimizations to better use domain specific information. We propose a new domain specific language, QIRAL, to generate parallel codes from high level formulations for Lattice QCD problems. QIRAL describes the formulation of the algorithms, of the matrices and preconditions used in this domain and generalizes languages such as SPIRAL used in auto-tuning library generator for signal processing applications. Lattice QCD applications require huge amount of processing power, on multinode, multi-core with GPUs. Simulation codes require to find new algorithms and efficient parallelization. So far, the difficulties for orchestrating parallelism efficiently hinder algorithmic exploration. The objective of QIRAL is to decouple algorithm exploration with parallelism description. Compiling QIRAL uses rewriting techniques for algorithm exploration, parallelization techniques for parallel code generation and potentially, runtime support to orchestrate this parallelism. Results of this work have been published in [12].

Following this effort, and through the combined analysis of the code behavior, at compile time and at runtime, MAQAO can then help users to better pinpoint and quantify performance issues in OpenMP codes, find load imbalance between threads, size of working sets, false sharing situations... We proposed in [22] to combine static and dynamic dependence analysis for the detection of vectorization opportunities. MAQAO then estimates the potential gain that could be reached through vectorization and identifies the required code transformations, either by changing loop control or data layout.

Aside from greedily invading all these new cores, demanding HPC applications now throw excited glances at the appealing computing power left unharvested inside the graphical processing units (GPUs). A strong demand is arising from the application programmers to be given means to access this power without bearing an unaffordable burden on the portability side. Efforts have already been made by the community in this respect but the tools provided still are rather close to the hardware, if not to the metal. Hence, we decided to launch some investigations on addressing this issue. In particular, we have designed a programming environment named StarPU that enables the programmer to offload tasks onto such heterogeneous processing units and gives that programmer tools to fit tasks to processing units capability, tools to efficiently manage data moves to and from the offloading hardware and handles the scheduling of such tasks all in an abstracted, portable manner. The challenge here is to take into account the intricacies of all computation unit: not only the computation power is heterogeneous among the machine, but data transfers themselves have various behavior depending on the machine architecture and GPUs capabilities, and thus have to be taken into account to get the best performance from the underlying machine. As a consequence, StarPU not only pays attention to fully exploit each of the different computational resources at the same time by properly mapping tasks in a dynamic manner according to their computation power and task behavior by the means of scheduling policies, but it also provides a distributed shared-memory library that makes it possible to manipulate data across heterogeneous multicore architectures in a high-level fashion while being optimized according to the machine possibilities. In addition to this, the scheduling policy of StarPU has been modularized; this makes it easy to experiment with state of the art theoretical scheduling strategies. Last but not least, StarPU works over clusters, by extending the shared-memory view over the MPI communication library. This allows, with the same sequential-looking application source code, to tackle all architectures from small multicore systems to clusters of heterogeneous systems. We extended OpenCL capabilities by proposing to use, transparently, StarPU as an OpenCL device [23].

On complex multicore, heterogeneous architectures, memory accesses often correspond in HPC application to performance bottlenecks. Indeed, either the code is memory bound, and restructuring data layout in order to take advantage of any reuse or spacial locality is essential. If the architecture has different types of memory (such as GPU with texture caches for instance), the code should exploit their features. Or the code is compute bound and in this case, SIMD vectorization represents the key for achieving high performance. Data structures may need to be changed in order to allow the compiler to automatically vectorize, or to efficiently vectorize. performance may only

be reached only at the cost of data layout restructuration. In order to better optimize data layout and parallelization, we proposed performance model for the memory hierarchy [26], [12]. Compared to other existing models, this model takes into account the costs due to the coherence protocol, the contention and the capacity of caches. It is built on top of parallel micro-benchmark results and thus can adapt to a wide range of architectures, and it aggregates these benchmark results for large code performance prediction. This model has been applied with success to communications on shared memory machines [27]. For specific memory, we have explored the opportunities and benefits of data restructuration, in collaboration with CEA [31]. Finally, data restructuration for SIMDization have been explored through the performance tuning tool MAQAO [22].

Optimizing communications over high performance clusters and grids Using a large panel of mechanisms such as user-mode communications, zero-copy transactions and communication operation offload, the critical path in sending and receiving a packet over high speed networks has been drastically reduced over the years. Recent implementations of the MPI standard, which have been carefully designed to directly map *basic* point-to-point requests onto the underlying low-level interfaces, almost reach the same level of performance for very basic point-to-point messaging requests. However more complex requests such as non-contiguous messages are left mostly unattended, and even more so are the irregular and multiflow communication schemes. The intent of the work on our NEWMADELEINE communication engine, for instance, is to address this situation thoroughly. The NEWMADELEINE optimization layer delivers much better performance on *complex* communication schemes with negligible overhead on basic single packet point-to-point requests. Through Mad-MPI, our proof-of-concept implementation of a subset of the MPI API, we intend to show that MPI applications can also benefit from the NEWMADELEINE communication engine.

The increasing number of cores in cluster nodes also raises the importance of intra-node communication. Our KNEM software module aims at offering optimized communication strategies for this special case and let the above MPI implementations benefit from dedicated models depending on process placement and hardware characteristics.

Moreover, the convergence between specialized high-speed networks and traditional ETHERNET networks leads to the need to adapt former software and hardware innovations to new message-passing stacks. Our work on the OPEN-MX software is carried out in this context.

Regarding larger scale configurations (clusters of clusters, grids), we intend to propose new models, principles and mechanisms that should allow to combine communication handling, threads scheduling and I/O event monitoring on such architectures, both in a portable and efficient way. We particularly intend to study the introduction of new runtime system functionalities to ease the development of code-coupling distributed applications, while minimizing their unavoidable negative impact on the application performance.

Integrating Communications and Multithreading Asynchronism is becoming ubiquitous in modern communication runtimes. Complex optimizations based on online analysis of the communication schemes and on the de-coupling of the request submission vs processing. Flow multiplexing or transparent heterogeneous networking also imply an active role of the runtime system request submit and process. And communication overlap as well as reactiveness are critical. Since network request cost is in the order of magnitude of several thousands CPU cycles at least, independent computations should not get blocked by an ongoing network transaction. This is even more true with the increasingly dense SMP, multicore, SMT architectures where many computing units share a few NICs. Since portability is one of the most important requirements for communication runtime systems, the usual approach to implement asynchronous processing is to use threads (such as Posix threads). Popular communication runtimes indeed are starting to make use of threads internally and also allow applications to also be multithreaded. Low level communication libraries also make use of multithreading. Such an introduction of threads inside communication subsystems is not going without troubles however. The fact that multithreading is still usually optional with these runtimes is symptomatic of the difficulty to get the benefits of multithreading in the context of networking without suffering from the potential drawbacks. We advocate the importance of the cooperation between

the asynchronous event management code and the thread scheduling code in order to avoid such disadvantages. We intend to propose a framework for symbiotically combining both approaches inside a new generic I/O event manager.

Moreover, the design of distributed parallel code, integrating both MPI and OpenMP, is complex and error-prone. Deadlock situations may arise and are difficult to detect. We proposed an original approach, based on static (compile-time) analysis and runtime verification in order to detect deadlock situation but also to pinpoint the cause of such deadlock [28], [15].

SCALE Team

3. Research Program

3.1. Safely and easily programming large-scale distributed applications

Our first objective is to provide a programming model for multi-level parallelism adapted to the programming of both multi-core level parallelism, and of large-scale distributed systems. Experience shows that achieving efficient parallelism at different levels with a single abstraction is difficult, however we will take particular care to provide a set of abstractions that are well integrated and form a safe and efficient global programming model. This programming model should also provide particular support for adaptation and dynamicity of applications.

3.1.1. Basic model

The main programming abstraction we have started to explore is multi-active object. This is a major change in the programming model since we remove the strongest constraint of active objects: their mono-threaded nature. Mono-threaded active objects bring powerful properties to our programming model, but also several limitations, including inefficiency on multicore machines, and deadlocks difficult to avoid. Thus, our objective here is to gain efficiency and expressiveness while maintaining as many properties of the original ASP calculus as possible, including ease of programming. Multi-active objects is a valuable alternative to the languages *à la* Creol/JCobox/ABS, as it is more efficient and potentially easier to program. This programming model better unifies the notions of concurrent programming and distributed programming, it is thus a crucial building block of our unified programming model.

It is also important to study related concurrency paradigms. Indeed, multi-active objects will not provide a complete solution to low-level concurrency; for this we should study the relation and the integration with other models for concurrency control (different programming languages, transactional memory models, ...).

Even if a first version of the language is available, further developments are necessary. In particular, the formal study of its properties is still an open subject. This formalisation is crucial in order to guarantee the correctness of the programming model. We have a good informal vision of the properties of the language but proving and formalising them is challenging due to the richness of the language.

3.1.2. Higher-level features

Multi-active objects should provide a good programming model integrating fine grain parallelism with large-scale distribution. We also think that the programming abstractions existing at the lower levels should nicely be integrated and interact with coarser-grain composition languages, in order to provide a unified programming model for multi-level parallelism. We think that it is also crucial, for the practical usability of the language to *design higher-level synchronisation primitives*. Indeed, a good basic programming language is not sufficient for its adoption in a real setting. Richer synchronisation primitives are needed to simply write complex interactions between entities running in parallel. The coexistence of several levels of parallelism will trigger the need for new primitives synchronising those several levels. Then the implementation of those primitives will require the design of new communication protocols that should themselves be formalised and verified.

One of the objectives of SCALE is also to provide frameworks for composing applications made of interacting distributed entities. The principle here would be to build basic composing blocks, typically made of a few multi-active objects, and then to compose an application made of these blocks using a coarser grain composition, like software components. What is particularly interesting is that we realised that software components also provide a component abstraction for reasoning on (compositional) program verification, or on autonomic adaptation of software and that active objects provide programming abstractions that fit well with software components. In the last years, the researchers of SCALE proposed GCM, a component model adapted to distribution and autonomic behaviour. We will reuse these results and adapt them. An even more challenging perspective consists in the use of component models for specifying discrete-event based simulations made up of different concerns; this will be a strong connection point between objective 1 and 3.

Finally, there still exists a gap between traditional programming languages like multi-active objects and coarse-grain composition languages like map-reduce paradigm. We want to investigate the interactions between these multiple layers of parallelism and provide a unified programming model.

3.1.3. Reliability of distributed applications

From the rigorous formalisation of the programming model(s), to the (assisted) proofs of essential properties, the use of model-checking-based methods for validating early system development, the range of formal method tools we use is quite large but the members of the teams are knowledgeable in those aspects. We also expect to provide tools to the programmers based on MDE approaches (with code-generation). While we might provide isolated contribution to theoretical domains, our objective is more to contribute to the applicability of formal methods in real development and runtime environments. We shall adapt our behavioural specification and verification techniques to the concurrency allowed in multi-active objects. Being able to ensure safety of multi-active objects will be a crucial tool, especially because those objects will be less easy to program than mono-threaded active objects.

Our experience has shown that model-checking methods, even when combining advanced abstraction techniques, state-of-the-art state-space representation, compositional approaches, and large-scale distributed model-checking engines, is (barely) able to master "middle-size" component systems using one complex interaction pattern (many-to-many communications), and/or a simple set of reconfiguration. If we want to be able to model complex features of distributed systems, and to reason on autonomic software components, verification techniques must scale. We strongly believe that further scalability will come from combination of theorem-proving and model-checking approaches. In a first step, theorem-proving can be used to prove generic properties of the model, that can be used to build smaller behavioural models, and reduce the model-checking complexity (reducing the model size, using symmetry properties, etc.). In a second step, we will use model-checking techniques on symbolic models that will rely on theorem proving for discharging proof obligations.

3.2. Easily, safely and efficiently running large-scale distributed applications

Concerning runtime aspect, a first necessary step is to to provide a runtime that can run efficiently the application written using the programming model described in objective 1. The proposed runtime environment will rely on commodity hosting platforms such as testbeds or clouds for being able to deploy and control, on demand, the necessary software stacks that will host the different applications components. The ProActive platform will be used as a basis that we will extend. Apart from autonomic adaptation aspects and their proof of correctness, we do not think that any new major research challenges will be solved here. However it is crucial to perform the necessary developments in order to show the practical effectiveness of our approach, and to provide a convenient and adaptable runtime to run the applications developed in the third objective about application domains.

3.2.1. Mapping and deploying virtual machines

The design of a cloud native application must follow established conventions. Among other things, true elasticity requires stateless components, load balancers, and queuing systems. The developer must also establish, with the cloud provider, the Service Level Agreements (SLAs) that state the quality of services to offer. For example, the amount of resources to allocate, the availability rate or possible placement criteria. In a private cloud, when the SLA implementation is not available, the application developer might be interested in implementing its own. Each developer must then master cloud architecture patterns and design his/her code accordingly. For example, he must be sure there is no single point of failures, that every elastic components is stateless that the balancing algorithms do not loose requests upon slave arrival and departure or the messaging protocol inside the queuing system is compatible with his/her usage. To implement a SLA enforcement algorithm, the developer must also master several families of combinatorial problems such as assignment and task scheduling, and ensure that the code fits the many possible situations. For example, he must consider the implication of every possible VM state on the resource consumption. As a result, the development and the deployment of performant cloud application require excessive skills for the developers.

The first original aspect we will push in this domain is related to safety and verification. It is established that OS kernels are critical softwares and many works proposed design to make them trustable through kernels and driver verifications. The VM scheduler is the new OS kernel but despite the economical damages a bug can cause, no one currently proposes any solution other than unit testing to improve the situation. As a result, production clouds currently run defective implementations. To address this critical situation we propose to formalise the specifications of VM scheduling primitives. Any developer should be able to specify his/her primitives. To fit their limited expertise in existing formal language, we will investigate for a domain specific language. This language will be used to prove the specified primitives with respect to the scheduler invariants. Second, it will make possible to generate the code of critical scheduler components. Typically the SLA enforcement algorithms. Third, the language will be used to assist at debugging legacy code and exhibit implementation bugs. Fabien Hermenier is already developing a language for specifying constraints for our research prototype VM Scheduler *BtrPlace*. *SafePlace* will be the name of the verification platform, we started its design and development in 2014.

The second challenge in this domain is to investigate the relation between programming languages, VM placement algorithms, allocation of resources, elasticity and adaptation concerns. The goal here is to enable the programmer to easily write and deploy scalable cloud applications by hiding with our programming model, the mechanisms the developer currently has to deal with explicitly today. This includes among other things to make transparent the notion of elastic components, elasticity rules, load balancing, or message queuing.

3.2.2. Debugging and fault-tolerance

We also aim at contributing to aspects that usually belong to pure distributed systems, generally from an algorithmic perspective. Indeed, we think that the approach we advocate is particularly interesting to bring new ideas to these research domains because of the interconnection between language semantics, protocols, and middleware. Typically, the knowledge we have on the programming model and on the behaviour of programs should help us provide dedicated debuggers and fault-tolerance protocols.

In fact some research has already been conducted in those domains, especially on reversible debuggers that allow the navigation inside a concurrent execution, doing forward and backward steps [0]. We think that those related works show that our approach is both relevant and timely. Moreover, little has been done for systems based on actors and active objects. The contribution we aim here is to provide debuggers able to better observe, introspect, and replay distributed executions. Such a tool will be of invaluable help to the programmer. Of course we will rely on existing tool for the local debugging and focus on the distributed aspects.

[0] Causal-Consistent Reversible Debugging. Elena Giachino, Ivan Lanese, and Claudio Antares Mezzina. *FASE 2014*.

SOCRATE Project-Team

3. Research Program

3.1. Research Axes

In order to keep young researchers in an environment close to their background, we have structured the team along the three research axes related to the three main scientific domains spanned by Socrate. However, we insist that a *major objective* of the Socrate team is to *motivate the collaborative research between these axes*, this point is specifically detailed in section 3.5 . The first one is entitled "Flexible Radio Front-End" and will study new radio front-end research challenges brought up by the arrival of MIMO technologies, and reconfigurable front-ends. The second one, entitled "Agile Radio Resource Sharing", will study how to couple the self-adaptive and distributed signal processing algorithms to cope with the multi-scale dynamics found in cognitive radio systems. The last research axis, entitled "Software Radio Programming Models" is dedicated to embedded software issues related to programming the physical protocols layer on these software radio machines. Figure 3 illustrates the three regions of a transceiver corresponding to the three Socrate axes.

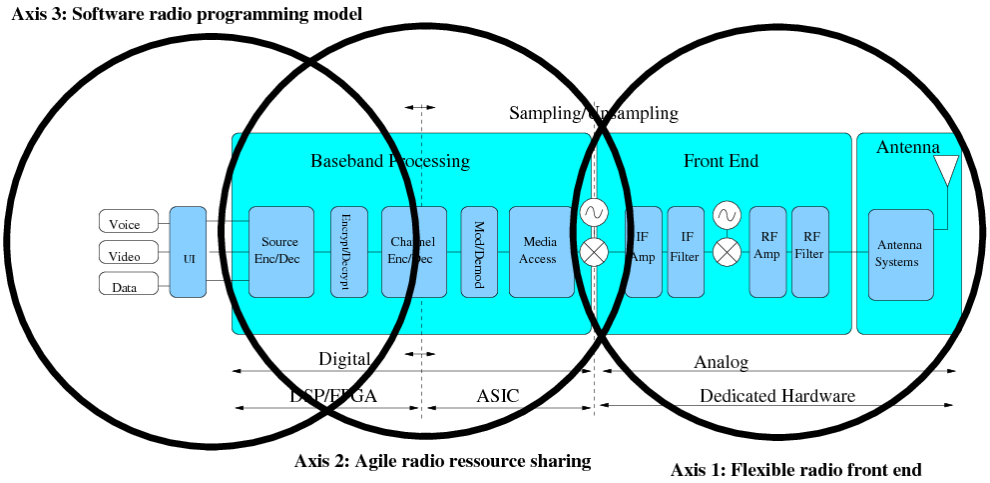


Figure 3. Center of interest for each of the three Socrate research axes with respect to a generic software radio terminal.

3.2. Flexible Radio Front-End

Participants: Guillaume Villemaud, Florin Hutu.

This axis mainly deals with the radio front-end of software radio terminals (right of Fig 3). In order to ensure a high flexibility in a global wireless network, each node is expected to offer as many degrees of freedom as possible. For instance, the choice of the most appropriate communication resource (frequency channel, spreading code, time slot,...), the interface standard or the type of antenna are possible degrees of freedom. The *multi-** paradigm denotes a highly flexible terminal composed of several antennas providing MIMO features to enhance the radio link quality, which is able to deal with several radio standards to offer interoperability and efficient relaying, and can provide multi-channel capability to optimize spectral reuse. On the other hand, increasing degrees of freedom can also increase the global energy consumption, therefore for energy-limited terminals a different approach has to be defined.

In this research axis, we expect to demonstrate optimization of flexible radio front-end by fine grain simulations, and also by the design of home made prototypes. Of course, studying all the components deeply would not be possible given the size of the team, we are currently not working in new technologies for DAC/ADC and power amplifiers which are currently studied by hardware oriented teams. The purpose of this axis is to build system level simulation taking into account the state of the art of each key component.

3.3. Agile Radio Resource Sharing

Participants: Jean-Marie Gorce, Claire Goursaud, Nikolai Lebedev, Perlaza Samir, Leonardo Sampaio-Cardoso.

The second research axis is dealing with the resource sharing problem between uncoordinated nodes but using the same (wide) frequency band. The agility represents the fact that the nodes may adapt their transmission protocol to the actual radio environment. Two features are fundamental to make the nodes agile : the first one is related to the signal processing capabilites of the software radio devices (middle circle in Fig 3), including modulation, coding, interference cancelling, sensing... The set of all available processing capabilites offers the degrees of freedom of the system. Note how this aspect relies on the two other research axes: radio front-end and radio programming.

But having processing capabilities is not enough for agility. The second feature for agility is the decision process, i.e. how a node can select its transmission mode. This decision process is complex because the appropriateness of a decision depends on the decisions taken by other nodes sharing the same radio environment. This problem needs distributed algorithms, which ensure stable and efficient solutions for a fair coexistence.

Beyond coexistence, the last decade saw a tremendous interest in cooperative techniques that let the nodes do more than coexisting. Of course, cooperation techniques at the networking or MAC layers for nodes implementing the same radio standard are well-known, especially for mobile ad-hoc networks, but cooperative techniques for SDR nodes at the PHY layer are still really challenging. The corresponding paradigm is the one of opportunistic cooperation, let us say *on-the-fly*, further implemented in a distributed manner.

We propose to structure our research into three directions. The two first directions are related to algorithmic developments, respectively for radio resource sharing and for cooperative techniques. The third direction takes another point of view and aims at evaluating theoretical bounds for different network scenarios using Network Information Theory.

The second research axis is dealing with multi-user communications focusing on resource sharing between uncoordinated nodes but using the same spectral resources. The agility relies on the nodes capability to adapt their transmission protocol to the actual radio environment. Centralized and decentralized approaches are investigated and the group is targeting fundamental limits as well as feasible and even practical implementations.

To make agile radio resource sharing a reality, two research directions are investigated. The first one aims at increasing the signal processing capabilities of software radio devices (middle circle in Fig 3), including modulation, coding, interference cancelation, sensing. The objective is to broaden the set of available processing capabilities thus offering more degrees of freedom. Note how this aspect relies on the two other research axes: radio front-end and radio programming.

Processing capabilities is not enough for agility. The second research direction concerns the decision process, i.e. how a node can select its transmission mode. This decision process is complex because the appropriateness of a decision depends on the decisions taken by other nodes sharing the same radio environment. In some cases, centralized solutions are possible but distributed algorithms are often required. Therefore, the target is to find distributed solutions ensuring stability, efficiency and fairness. Beyond coexistence, the last decade saw a tremendous interest in cooperative techniques that let the nodes do more than coexisting. Of course, cooperation techniques at the networking or MAC layers for nodes implementing the same radio standard are well-known, especially for mobile ad-hoc networks, but cooperative techniques for SDR nodes at the PHY layer are still challenging. The corresponding paradigm is referred to as opportunistic cooperative transmissions. We structure our research into three directions:

- Establishing theoretical limits of cooperative wireless networks in the network information theory framework.
- Designing coding and signal processing techniques for optimal transmissions (e.g. interference alignment).
- Developing distributed mechanisms for distributed decision at layer 1 and 2, using game theory, consensus and graph modeling.

3.4. Software Radio Programming Model

Participants: Tanguy Risset, Kevin Marquet, Guillaume Salagnac, Florent de Dinechin.

Finally the third research axis is concerned with software aspect of the software radio terminal (left of Fig 3). We have currently two actions in this axis, the first one concerns the programming issues in software defined radio devices, the second one focusses on low power devices: how can they be adapted to integrate some reconfigurability.

The expected contributions of Socrate in this research axis are :

- The design and implementation of a "middleware for SDR", probably based on a Virtual Machine.
- Prototype implementations of novel software radio systems, using chips from Leti and/or Lyrtech software radio boards [0].
- Development of a *smart node*: a low-power Software-Defined Radio node adapted to WSN applications.
- Methodology clues and programming tools to program all these prototypes.

3.5. Inter-Axes collaboration

Innovative results come from collaborations between the three axes. To highlight the fact that this team structure does not limit the ability of inter-axes collaborations between Socrate members, we list below the *on-going* research actions that *already* involve actors from two or more axes, this is also represented on Fig 4 .

- *Optimizing network capacity of very large scale networks*. 2 Phds started in October/November 2011 with Guillaume Villemaud (axis 1) and Claire Goursaud (axis 2), respectively.
- *SDR for sensor networks*. A PhD started in 2012 in collaboration with FT R&D, involving people from axis 3 (Guillaume Salagnac, Tanguy Risset) and axis 1 (Guillaume Villemaud).
- *CorteXlab*. The 3 axes also collaborate on the design and the development of CorteXlab.
- *body area networks applications*. Axis 2 and axis 3 collaborate on the development of body area networks applications in the framework of the FUI Smacs project. Jean-Marie Gorce and Tanguy Risset co-advised Matthieu Lauzier.
- *Wiplan and NS3*. The MobiSim ADT involves Guillaume Villemaud (axis 1) and Jean-Marie Gorce (axis 2).
- *Resource allocation and architecture of low power multi-band front-end*. The EconHome project involves people from axis 2 (Jean-Marie Gorce,Nikolai Lebedev) and axis 1 (Florin Hutu). 1 Phd started in 2011.
- *Virtual machine for SDR*. In collaboration with CEA, a PhD started in October 2011, involving people from axis 3 (Tanguy Risset, Kevin Marquet) and Leti's engineers closer to axis 2.
- *Relay strategy for cognitive radio*. Guillaume Villemaud and Tanguy Risset were together advisers of Cedric Levy-Bencheton PhD Thesis (defense last June).

Finally, we insist on the fact that the *FIT project* will involve each member of Socrate and will provide many more opportunities to perform cross layer SDR experimentations. FIT is already federating all members of the Socrate team.

[0] Lyrtech (http://www.lyrtech.com) designs and sells radio card receivers with multiple antennas offering the possibility to implement a complete communication stack

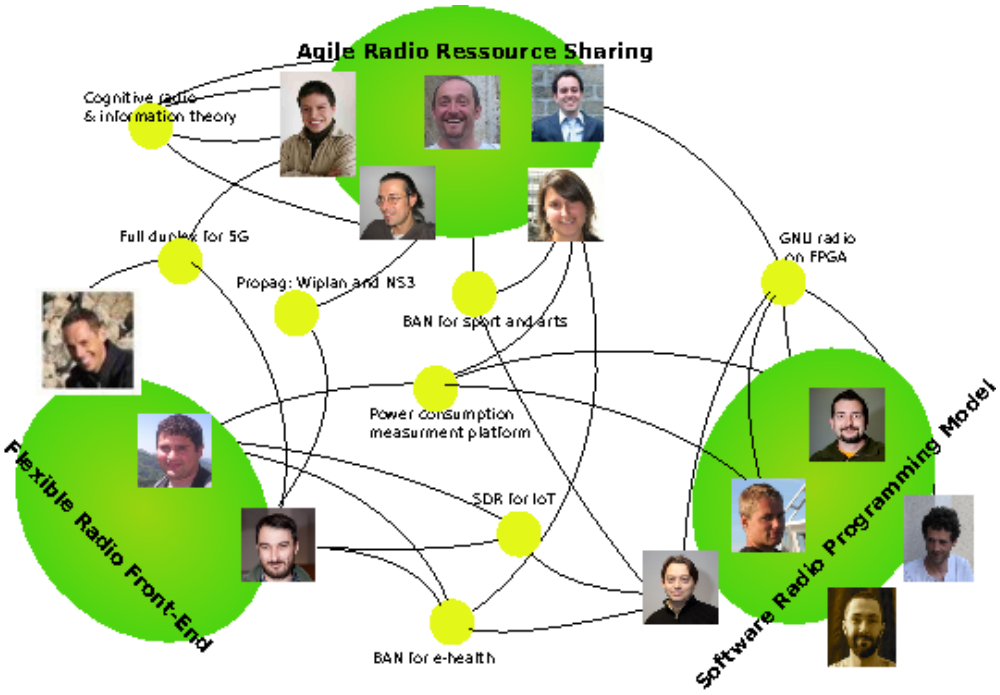


Figure 4. Inter-Axis Collaboration in Socrate: we expect innovative results to come from this pluri-disciplinary research

SPIRALS Team

3. Research Program

3.1. Introduction

Our research program on self-adaptive software targets two key properties that are detailed in the remainder of this section: self-healing and self-optimization.

3.2. Objective #1: Self-healing - Mining software artifacts to automatically evolve systems

Software systems are under the pressure of changes all along their lifecycle. Agile development blurs the frontier between design and execution and requires constant adaptation. The size of systems (millions of lines of code) multiplies the number of bugs by the same order of magnitude. More and more systems, such as sensor network devices, live in "surviving" mode, in the sense that they are neither rebootable nor upgradable.

Software bugs are hidden in source code and show up at development-time, testing-time or worse, once deployed in production. Except for very specific application domains where formal proofs are achievable, bugs can not be eradicated. As an order of magnitude, on 16 Dec 2011, the Eclipse bug repository contains 366,922 bug reports. Software engineers and developers work on bug fixing on a daily basis. Not all developers spend the same time on bug fixing. In large companies, this is sometimes a full-time role to manage bugs, often referred to as Quality Assurance (QA) software engineers. Also, not all bugs are equal, some bugs are analyzed and fixed within minutes, others may take months to be solved [123].

In terms of research, this means that: (i) one needs means to automatically adapt the design of the software system through automated refactoring and API extraction, (ii) one needs approaches to automate the process of adapting source code in order to fix certain bugs, (iii) one needs to revisit the notion of error-handling so that instead of crashing in presence of errors, software adapts itself to continue with its execution, e.g., in degraded mode.

There is no one-size-fits-all solution for each of these points. However, we think that novel solutions can be found by using **data mining and machine learning techniques tailored for software engineering** [124]. This body of research consists of mining some knowledge about a software system by analyzing the source code, the version control systems, the execution traces, documentation and all kinds of software development and execution artifacts in general. This knowledge is then used within recommendation systems for software development, auditing tools, runtime monitors, frameworks for resilient computing, etc.

The novelty of our approach consists of using and tailoring data mining techniques for analyzing software artifacts (source code, execution traces) in order to achieve the **next level of automated adaptation** (*e.g.*, automated bug fixing). Technically, we plan to mix unsupervised statistical learning techniques (*e.g.* frequent item set mining) and supervised ones (*e.g.* training classifiers such as decision trees). This research is currently not being performed by data mining research teams since it requires a high level of domain expertise in software engineering, while software engineering researchers can use off-the-shelf data mining libraries, such as Weka [98].

We now detail the two directions that we propose to follow to achieve this objective.

3.2.1. Learning from software history how to design software and fix bugs

The first direction is about mining techniques in software repositories (*e.g.*, CVS, SVN, Git). Best practices can be extracted by data mining source code and the version control history of existing software systems. The design and code of expert developers significantly vary from the artifacts of novice developers. We will learn to differentiate those design characteristics by comparing different code bases, and by observing the semantic refactoring actions from version control history. Those design rules can then feed the test-develop-refactor constant adaptation cycle of agile development.

Fault localization of bugs reported in bug repositories. We will build a solid foundation on empirical knowledge about bugs reported in bug repository. We will perform an empirical study on a set of representative bug repositories to identify classes of bugs and patterns of bug data. For this, we will build a tool to browse and annotate bug reports. Browsing will be helped with two kinds of indexing: first, the tool will index all textual artifacts for each bug report; second it will index the semantic information that is not present by default in bug management software (*i.e.*, "contains a stacktrace"). Both indexes will be used to find particular subsets of bug reports, for instance "all bugs mentioning invariants and containing a stacktrace". Note that queries with this kind of complexity and higher are mostly not possible with the state-of-the-art of bug management software. Then, analysts will use annotation features to annotate bug reports. The main outcome of the empirical study will be the identification of classes of bugs that are appropriate for automated localization. Then, we will run machine learning algorithms to identify the latent links between the bug report content and source code features. Those algorithms would use as training data the existing traceability links between bug reports and source code modifications from version control systems. We will start by using decision trees since they produce a model that is explicit and understandable by expert developers. Depending on the results, other machine learning algorithms will be used. The resulting system will be able to locate elements in source code related to a certain bug report with a certain confidence.

Automated bug fix generation with search-based techniques. Once a location in code is identified as being the cause of the bug, we can try to automatically find a potential fix. We envision different techniques: (1) infer fixes from existing contracts and specifications that are violated; (2) infer fixes from the software behavior specified as a test suite; (3) try different fix types one-by-one from a list of identified bug fix patterns; (4) search fixes in a fix space that consists of combinations of atomic bug fixes. Techniques 1 and 2 are explored in [91] and [122]. We will focus on the latter techniques. To identify bug fix patterns and atomic bug fixes, we will perform a large-scale empirical study on software changes (also known as changesets when referring to changes across multiple files). We will develop tools to navigate, query and annotate changesets in a version control system. Then, a grounded theory will be built to master the nature of fixes. Eventually, we will decompose change sets in atomic actions using clustering on changeset actions. We will then use this body of empirical knowledge to feed search-based algorithms (*e.g.* genetic algorithms) that will look for meaningful fixes in a large fix space. To sum up, our research on automated bug fixing will try not only to point to source code locations responsible of a bug, but to search for code patterns and snippets that may constitute the skeleton of a valid patch. Ultimately, a blend of expert heuristics and learned rules will be able to produce valid source code that can be validated by developers and committed to the code base.

3.2.2. Run-time self-healing

The second proposed research direction is about inventing a self-healing capability at run-time. This is complementary to the previous objective that mainly deals with development time issues. We will achieve this in two steps. First, we want to define frameworks for resilient software systems. Those frameworks will help to maintain the execution even in the presence of bugs, *i.e.* to let the system survive. As exposed below, this may mean for example to switch to some degraded modes. Next, we want to go a step further and to define solutions for automated runtime repair, that is, not simply compensating the erroneous behavior, but also determining the correct repair actions and applying them at run-time.

Mining best effort values. A well-known principle of software engineering is the "fail-fast" principle. In a nutshell, it states that as soon as something goes wrong, software should stop the execution before entering incorrect states. This is fine when a human user is in the loop, capable of understanding the error or at least rebooting the system. However, the notion of "failure-oblivious computing" [112] shows that in certain domains, software should run in a resilient mode (*i.e.* capable of recovering from errors) and/or best-effort mode (*i.e.* a slightly imprecise computation is better than stopping). Hence, we plan to investigate data mining techniques in order to learn best-effort values from past executions (*i.e.* somehow learning what is a correct state, or the opposite what is not a completely incorrect state). This knowledge will then be used to adapt the software state and flow in order to mitigate the error consequences, the exact opposite of fail-fast for systems with long-running cycles.

Embedding search based algorithms at runtime. Harman recently described the field of search-based software engineering [99]. We think that certain search based approaches can be embedded at runtime with the goal of automatically finding solutions that avoid crashing. We will create software infrastructures that allow automatically detecting and repairing faults at run-time. The methodology for achieving this task is based on three points: (1) empirical study of runtime faults; (2) learning approaches to characterize runtime faults; (3) learning algorithms to produce valid changes to the software runtime state. An empirical study will be performed to analyze those bug reports that are associated with runtime information (*e.g.* core dumps or stacktraces). After this empirical study, we will create a system that learns on previous repairs how to produce small changes that solve standard runtime bugs (*e.g.* adding an array bound check to throw a handled domain exception rather than a spurious language exception). To achieve this task, component models will be used to (1) encapsulate the monitoring and reparation meta-programs in appropriate components and (2) support runtime code modification using scripting, reflective or bytecode generation techniques.

3.3. Objective #2: Self-optimization - Sharing runtime behaviors to continuously adapt software

Complex distributed systems have to seamlessly adapt to a wide variety of deployment targets. This is due to the fact that developers cannot anticipate all the runtime conditions under which these systems are immersed. A major challenge for these software systems is to develop their capability to continuously reason about themselves and to take appropriate decisions and actions on the optimizations they can apply to improve themselves. This challenge encompasses research contributions in different areas, from environmental monitoring to realtime symptoms diagnosis, to automated decision making. The variety of distributed systems, the number of optimization parameters, and the complexity of decisions often resign the practitioners to design monolithic and static middleware solutions. However, it is now globally acknowledged that the development of dedicated building blocks does not contribute to the adoption of sustainable solutions. This is confirmed by the scale of actual distributed systems, which can—for example—connect several thousands of devices to a set of services hosted in the Cloud. In such a context, the lack of support for smart behaviours at different levels of the systems can inevitably lead to its instability or its unavailability. In June 2012, an outage of Amazon's Elastic Compute Cloud in North Virginia has taken down Netflix, Pinterest, and Instagram services. During hours, all these services failed to satisfy their millions of customers due to the lack of integration of a self-optimization mechanism going beyond the boundaries of Amazon.

The research contributions we envision within this area will therefore be organized as a reference model for engineering **self-optimized distributed systems** autonomously driven by *adaptive feedback control loops*, which will automatically enlarge their scope to cope with the complexity of the decisions to be taken. This solution introduces a multi-scale approach, which first privileges local and fast decisions to ensure the homeostasis [0] property of a single node, and then progressively propagates symptoms in the network in order to reason on a longer term and a larger number of nodes. Ultimately, domain experts and software developers can be automatically involved in the decision process if the system fails to find a satisfying solution. The research program for this objective will therefore focus on the study of mechanisms for **monitoring, taking decisions, and automatically reconfiguring software at runtime and at various scales**. As stated in the self-healing objective, we believe that there is no one-size-fits-all mechanism that can span all the scales of the system. We will therefore study and identify an optimal composition of various adaptation mechanisms in order to produce long-living software systems.

The novelty of this objective is to exploit the wisdom of crowds to define new middleware solutions that are able to continuously adapt software deployed in the wild. We intend to demonstrate the applicability of this approach to distributed systems that are deployed from mobile phones to cloud infrastructures. The key scientific challenges to address can be summarized as follows: *How does software behave once deployed in the wild? Is it possible to automatically infer the quality of experience, as it is perceived by users? Can the*

[0] Homeostasis is the property of a system that regulates its internal environment and tends to maintain a stable, relatively constant condition of properties [Wikipedia].

runtime optimizations be shared across a wide variety of software? How optimizations can be safely operated on large populations of software instances?

The remainder of this section further elaborates on the opportunities that can be considered within the frame of this objective.

3.3.1. Monitoring software in the wild

Once deployed, developers are generally no longer aware of how their software behave. Even if they heavily use testbeds and benchmarks during the development phase, they mostly rely on the bugs explicitly reported by users to monitor the efficiency of their applications. However, it has been shown that contextual artifacts collected at runtime can help to understand performance leaks and optimize the resilience of software systems [125]. Monitoring and understanding the context of software at runtime therefore represent the first building block of this research challenge. Practically, we intend to investigate crowdsensing approaches, to smartly collect and process runtime metrics (*e.g.*, request throughput, energy consumption, user context). Crowdsensing can be seen as a specific kind of crowdsourcing activity, which refers to the capability of lifting a (large) diffuse group of participants to delegate the task of retrieving trustable data from the field. In particular, crowdsensing covers not only *participatory sensing* to involve the user in the sensing task (*e.g.*, surveys), but also *opportunistic sensing* to exploit mobile sensors carried by the user (*e.g.*, smartphones).

While reported metrics generally enclose raw data, the monitoring layer intends to produce meaningful indicators like the *Quality of Experience* (QoE) perceived by users. This QoE reflects representative symptoms of software requiring to trigger appropriate decisions in order to improve its efficiency. To diagnose these symptoms, the system has to process a huge variety of data including runtime metrics, but also history of logs to explore the sources of the reported problems and identify opportunities for optimizations. The technics we envision at this level encompass machine learning, principal component analysis, and fuzzy logic [111] to provide enriched information to the decision level.

3.3.2. Collaborative decision-making approaches

Beyond the symptoms analysis, decisions should be taken in order to improve the *Quality of Service* (QoS). In our opinion, collaborative approaches represent a promising solution to effectively converge towards the most appropriate optimization to apply for a given symptom. In particular, we believe that exploiting the wisdom of the crowd can help the software to optimize itself by sharing its experience with other software instances exhibiting similar symptoms. The intuition here is that the body of knowledge that supports the optimization process cannot be specific to a single software instance as this would restrain the opportunities for improving the quality and the performance of applications. Rather, we think that any software instance can learn from the experience of others.

With regard to the state-of-the-art, we believe that a multi-levels decision infrastructure, inspired from distributed systems like Spotify [95], can be used to build a decentralized decision-making algorithm involving the surrounding peers before requesting a decision to be taken by more central control entity. In the context of collaborative decision-making, peer-based approaches therefore consist in quickly reaching a consensus on the decision to be adopted by a majority of software instances. Software instances can share their knowledge through a micro-economic model [89], that would weight the recommandations of experienced instances, assuming their age reflects an optimal configuration.

Beyond the peer level, the adoption of algorithms inspired from evolutionary computations, such as genetic programming, at an upper level of decision can offer an opportunity to test and compare several alternative decisions for a given symptom and to observe how does the crowd of applications evolves. By introducing some diversity within this population of applications, some instances will not only provide a satisfying QoS, but will also become naturally resilient to unforeseen situations.

3.3.3. Smart reconfigurations in the large

Any decision taken by the crowd requires to propagate back to and then operated by the software instances. While simplest decisions tend to impact software instances located on a single host (*e.g.*, laptop, smartphone),

this process can also exhibit more complex reconfiguration scenarios that require the orchestration of various actions that have to be safely coordinated across a large number of hosts. While it is generally acknowledged that centralized approaches raise scalability issues, we think that self-optimization should investigate different reconfiguration strategies to propagate and apply the appropriate actions. The investigation of such strategies can be addressed in two steps: the consideration of *scalable data propagation protocols* and the identification of *smart reconfiguration mechanisms*.

With regard to the challenge of scalable data propagation protocols, we think that research opportunities encompass not only the exploitation of gossip-based protocols [94], but also the adoption of publish/subscribe abstractions [101] in order to decouple the decision process from the reconfiguration. The fundamental issue here is the definition of a communication substrate that can accomodate the propagation of decisions with relaxed properties, inspired by *Delay Tolerant Networks* (DTN), in order to reach weakly connected software instances. We believe that the adoption of asynchronous communication protocols can provide the sustainable foundations for addressing various execution environments including harsh environments, such as developing countries, which suffer from a partial connectivity to the network. Additionally, we are interested in developing the principle of *social networks of applications* in order to seamlessly group and organize software instances according to their similarities and acquaintances. The underlying idea is that grouping application instances can contribute to the identification of optimization profiles not only contributing to the monitoring layer, but also interested in similar reconfigurations. Social networks of applications can contribute to the anticipation of reconfigurations by exploiting the symptoms of similar applications to improve the performance of others before that problems actually happen.

With regard to the challenge of smart reconfiguration mechanisms, we are interested in building on our established experience of adaptive middleware [8] in order to investigate novel approaches to efficient application reconfigurations. In particular, we are interested in adopting seamless micro-updates and micro-reboot technics to provide in-situ reconfiguration of pieces of software. Additionally, the provision of safe and secured reconfiguration mechanisms is clearly a key issue that requires to be carefully addressed in order to avoid malicious exploitation of dynamic reconfiguration mechanisms against the software itself. In this area, although some reconfiguration mechanisms integrate transaction models [102], most of them are restricted to local reconfigurations, without providing any support for executing distributed reconfiguration transactions. Additionally, none of the approached published in the literature include security mechanisms to preserve from unauthorized or malicious reconfigurations.

TACOMA Team

3. Research Program

3.1. Using and Programming Context

The goal of ambient computing is to seamlessly merge virtual and real environments. A real environment is composed of objects from the physical world, e.g., people, places, machines. A virtual environment is any information system, e.g., the Web. The integration of these environments must permit people and their information systems to implicitly interact with their surrounding environment.

Ambient computing applications are able to evaluate the state of the real world through sensing technologies. This information can include the position of a person (caught with a localization system like GPS), the weather (captured using specialized sensors), etc. Sensing technologies enable applications to automatically update digital information about events or entities in the physical world. Further, interfaces can be used to act on the physical world based on information processed in the digital environment. For example, the windows of a car can be automatically closed when it is raining.

This real-world and virtual-world integration must permit people to implicitly interact with their surrounding environment. This means that manual device manipulation must be minimal since this constrains person mobility. In any case, the relative small size of personal devices can make them awkward to manipulate. In the near future, interaction must be possible without people being aware of the presence of neighbouring processors.

Information systems require tools to *capture* data in its physical environment, and then to *interpret*, or process, this data. A context denotes all information that is pertinent to a person-centric application. There are three classes of context information:

- The *digital context* defines all parameters related to the hardware and software configuration of the device. Examples include the presence (or absence) of a network, the available bandwidth, the connected peripherals (printer, screen), storage capacity, CPU power, available executables, etc.
- The *personal context* defines all parameters related to the identity, preferences and location of the person who owns the device. This context is important for deciding the type of information that a personal device needs to acquire at any given moment.
- The *physical context* relates to the person's environment; this includes climatic condition, noise level, luminosity, as well as date and time.

All three forms of context are fundamental to person-centric computing. Consider for instance a virtual museum guide service that is offered via a PDA. Each visitor has his own PDA that permits him to receive and visualise information about surrounding artworks. In this application, the *pertinent* context of the person is made up of the artworks situated near the person, the artworks that interest him as well as the degree of specialisation of the information, i.e., if the person is an art expert, he will desire more detail than the occasional museum visitor.

There are two approaches to organising data in a real to virtual world mapping: a so-called *logical* approach and a *physical* approach. The logical approach is the traditional way, and involves storing all data relevant to the physical world on a service platform such as a centralised database. Context information is sent to a person in response to a request containing the person's location co-ordinates and preferences. In the example of the virtual museum guide, a person's device transmits its location to the server, which replies with descriptions of neighbouring artworks.

The main drawbacks of this approach are scalability and complexity. Scalability is a problem since we are evolving towards a world with billions of embedded devices; complexity is a problem since the majority of physical objects are unrelated, and no management body can cater for the integration of their data into a service platform. Further, the model of the physical world must be up to date, so the more dynamic a system, the more updates are needed. The services platform quickly becomes a potential bottleneck if it must deliver services to all people.

The physical approach does not rely on a digital model of the physical world. The service is computed wherever the person is located. This is done by spreading data onto the devices in the physical environment; there are a sufficient number of embedded systems with wireless transceivers around to support this approach. Each device manages and stores the data of its associated object. In this way, data are physically linked to objects, and there is no need to update a positional database when physical objects move since the data *physically* moves with them.

With the physical approach, computations are done on the personal and available embedded devices. Devices interact when they are within communication range. The interactions constitute delivery of service to the person. Returning to the museum example, data is directly embedded in a painting's frame. When the visitor's guide meets (connects) to a painting's devices, it receives the information about the painting and displays it.

3.2. Coupled objects

Integrity checking is an important concern in many activities, both in the real world and in the information society. The basic purpose is to verify that a set of objects, parts, components, people remains the same along some activity or process, or remains consistent against a given property (such as a part count).

In the real world, it is a common step in logistic: objects to be transported are usually checked by the sender (for their conformance to the recipient expectation), and at arrival by the recipient. When a school get a group of children to a museum, people responsible for the children will regularly check that no one is missing. Yet another common example is to check for our personal belongings when leaving a place, to avoid lost. While important, these verification are tedious, vulnerable to human errors, and often forgotten.

Because of these vulnerabilities, problems arise: E-commerce clients sometimes receive incomplete packages, valuable and important objects (notebook computers, passports etc.) get lost in airports, planes, trains, hotels, etc. with sometimes dramatic consequences.

While there are very few automatic solutions to improve the situation in the real world, integrity checking in the computing world is a basic and widely used mechanism: magnetic and optical storage devices, network communications are all using checksums and error checking code to detect information corruption, to name a few.

The emergence of ubiquitous computing and the rapid penetration of RFID devices enable similar integrity checking solutions to work for physical objects. We introduced the concept of *coupled object*, which offers simple yet powerful mechanisms to check and ensure integrity properties for set of physical objects.

Essentially, coupled objects are a set of physical objects which defines a logical group. An important feature is that the group information is self contained on the objects which allow to verify group properties, such as completeness, only with the objects. Said it another way, the physical objects can be seen as fragments of a composite object. A trivial example could be a group made of a person, his jacket, his mobile phone, his passport and his cardholder.

The important feature of the concept are its distributed, autonomous and anonymous nature: it allows the design and implementation of pervasive security applications without any database tracking or centralized information system support. This is a significant advantage of this approach given the strong privacy issues that affect pervasive computing.

TYREX Project-Team

3. Research Program

3.1. Modeling

Modeling consists in capturing various aspects of document and data processing and communication in a unifying model. Our modeling research direction mainly focuses on three aspects.

The first aspect aims at reducing the impedance mismatch. The impedance mismatch refers to the complexity, difficulty and lack of performance induced by various web application layers which require the same piece of information to be represented and processed differently. The mismatch occurs because programming languages use different native data models from those used for documents in browsers and for storage in databases. This results in complex and multi-tier software architectures whose different layers are incompatible in nature. This, in turn, results in expensive, inefficient, and error-prone web development. For reducing the impedance mismatch, we will focus on the design of a unifying software stack and programming framework, backed by generic and solid logical foundations similar in spirit to the NoSQL approach.

The second aspect aims at harnessing heterogeneity. Web applications increasingly use diverse data models: ordered and unordered tree-like structures (such as XML), nested records and arrays (such as JSON), graphs like (e.g. RDF), and tables. Furthermore, these data models also involve a variety of languages for expressing constraints over data (e.g. XML schema, the well-founded RelaxNG, and RDFS to name just a few). We believe that this heterogeneity is here to stay and is likely to increase. These differences in representations imply loads of error-prone and costly conversions and transformations. Furthermore, some native formats (e.g. JSON) are diverted from a programming construct to a data exchange one. This often results in a loss of information and in errors that need to be tracked and corrected. In this context, it is important to seek methods for reducing risks of information loss during data transformation and exchange. For harnessing heterogeneity, we will focus on the integration of data models through unified formal semantics and in particular logical interpretation. This allows using the same programming language constructs on different data models. At the programming language level, this is similar to languages such as JSonIq for JSON and XML.

Finally, the third aspect aims at making applications and data more compositional. Most web programming technologies are currently limited from a compositional point of view. For example, tree grammars (like schema languages for XML) are monolithic in the sense that they require the full description of the considered structures, instead of allowing the assembly of smaller and reusable building blocks. More generally, this need is illustrated in the industry by the increasing development of W3C specifications organised in ad-hoc modules. So far, these various attempts have failed to provide an acceptable mechanism for composition. For example, HTML5 has been specified in a monolithic way despite the fact that it relies on several other existing specifications (such as HTML, SVG, SMIL, CSS, etc.). As a consequence, this translates into monolithic web applications, which makes their automated verification harder by making modular analyses more difficult. For making applications and data more compositional, we will focus on the design of modular schema and programming languages. For this purpose, we will notably rely on succinct yet expressive formalisms (like two-way logics, polymorphic types) that ease the process of expressing modular specifications.

One major scientific difficulty in this overall direction consists in taking into account the specificities of the web, which require new programming models and supporting theoretical tools that do not exist today.

3.2. Analysis, verification and optimization

This research direction aims at guaranteeing two different kinds of properties: safety and efficiency.

The first kind of properties concern safety of web applications. Software development was traditionally split between critical and non-critical software. Advanced (and costly) formal verification techniques were reserved to the former whereas non-critical software relied almost exclusively on testing, which only offers a "best-effort" guarantee (removes most bugs but some of them may not be detected). The central idea was that in a non-critical system, the damage a failure may create is not worth the cost of formal verification. However as web applications grow more pervasive in everyday life and gain momentum in corporates, various social organizations, and touch larger numbers of users, the potential cost of failure is increasing rapidly and significantly. Despite this fact, it is more obvious, in healthcare for instance, to qualify as a critical component a pacemaker than the hospital's information system. Of course, a failure of such a device would directly cause death, however a general failure of the hospital's information system may cause deaths as well and possibly even incur greater damages. In that sense, we can consider that web applications are becoming more and more critical. The growing dependency on the web as a tool, combined with the fact that some applications involve very large user bases, is becoming problematic as it seems to increase rapidly but silently. Some errors like crashes and confidential information leaks, if not discovered, can have massive effects and incur significant financial or reputation damage.

The second kind of properties concern efficiency of web applications. One particular characteristic of web programming languages is that they are essentially data-manipulation oriented. These manipulations rely on query and transformation languages whose performance is critical. This performance is very sensitive to data size and organization (constraints) and to the execution model (e.g. streaming evaluators). Static analysis can be used to optimize runtime performance by compile-time automated modification of the code (e.g. substitution of queries by more efficient ones). One major scientific difficulty here consists in dealing with problems close to the frontier of decidability, and therefore in finding useful trade-offs between programming ease, expressivity, complexity, succinctness, algorithmic techniques and effective implementations.

URBANET Team

3. Research Program

3.1. Capillary networks

The definition of Smart Cities is still constantly redefined and expanded so as to comprehensively describe the future of major urban areas. The Smart City concept mainly refers to granting efficiency and sustainability in densely populated metropolitan areas while enhancing citizens' life and protecting the environment. The Smart City vision can be primarily achieved by a clever integration of ICT in the urban tissue. Indeed, ICTs are enabling an evolution from the current duality between the "real world" and its digitalized counterpart to a continuum in which digital contents and applications are seamlessly interacting with classical infrastructures and services. The general philosophy of smart cities can also be seen as a paradigm shift combining the Internet of Things (IoT) and Machine-to-Machine (M2M) communication with a citizen-centric model, all together leveraging massive data collected by pervasive sensors, connected mobile or fixed devices, and social applications.

The fast expansion of urban digitalization yields new challenges that span from social issues to technical problems. Therefore, there is a significant joint effort by public authorities, academic research communities and industrial companies to understand and address these challenges. Within that context, the application layer, i.e., the novel services that ICT can bring to digital urban environments, have monopolized the attention. Lower-layer network architectures have gone instead quite overlooked. We believe that this might be a fatal error, since the communication network plays a critical role in supporting advanced services and ultimately in making the Smart City vision a reality. The UrbaNet project deals precisely with that aspect, and the study of network solutions for upcoming Smart Cities represents the core of our work.

Most network-related challenges along the road to real-world Smart Cities deal with efficient mobile data communication, both at the backbone and at the radio access levels. It is on the latter that the UrbaNet project is focused. More precisely, the scope of the project maps to that of capillary networks, an original concept we define next.

The capillary networking concept represents a unifying paradigm for wireless last-mile communication in smart cities. The term we use is reminiscent of the pervasive penetration of different technologies for wireless communication in future digital cities. Indeed, capillary networks represent the very last portion of the data distribution and collection network, bringing Internet connectivity to every endpoint of the urban tissue in the same exact way capillary blood vessels bring oxygen and collect carbon dioxide at tissues in the human body. Capillary networks inherit concepts from the self-configuring, autonomous, ad hoc networks so extensively studied in the past decade, but they do so in a holistic way. Specifically, this implies considering multiple technologies and applications at a time, and doing so by accounting for all the specificities of the urban environment.

3.2. Specific issues and new challenges of capillary networks

Capillary networks are not just a collection of independent wireless technologies that can be abstracted from the urban environment and/or studied separately. That approach has been in fact continued over the last decade, as technologies such as sensor, mesh, vehicular, opportunistic, and – generally speaking – M2M networks have been designed and evaluated in isolation and in presence of unrealistic mobility and physical layer, simplistic deployments, random traffic demands, impractical application use cases and non-existent business models. In addition, the physical context of the network has a significant impact on its performances and cannot be reduced to a simple random variable. Moreover, one of the main element of a network never appears in many studies: the user. To summarize, networks issues should be addressed from a user- and context-centric perspective.

Such abstractions and approximations were necessary for understanding the fundamentals of wireless network protocols. However, real world deployments have shown their limits. The finest protocols are often unreliable and hardly applicable to real contexts. That also partially explains the marginal impact of multi-hop wireless technologies on today's production market. Industrial solutions are mostly single-hop, complex to operate, and expensive to maintain.

In the UrbaNet project we consider the capillary network as an ensemble of strongly intertwined wireless networks that are expected to coexist and possibly co-operate in the context of arising digital cities. This has three major implications:

- Each technology contributing to the overall capillary network should not be studied apart. As a matter of fact, mobile devices integrate today a growing number of sensors (e.g., environment sensing, resource consumption metering, movement, health or pollution monitoring) and multiple radio interfaces (e.g., LTE, WiFi, ZigBee,. . .), and this is becoming a trend also in the case of privately owned cars, public transport vehicles, commercial fleets, and even city bikes. Similarly, access network sites tend to implement heterogeneous communication technologies so as to limit capital expenses. Enabling smart-cities needs a dense sensing of its activities, which cannot be achieved without multi-service sensor networks. Moreover, all these devices are expected to inter-operate so as to make the communication more sustainable and reliable. Thus, the technologies that build up the capillary network shall be studied as a whole in the future.
- The capillary network paradigm necessarily accounts for actual urban mobility flows, city land-use layouts, metropolitan deployment constraints, and expected activity of the citizens. Often, these specificities do not arise from purely networking features, but relate to the study of city topologies and road layouts, social acceptability, transportation systems, energy management, or urban economics. Therefore, addressing capillary network scenarios cannot but rely on strong multidisciplinary interactions.
- Digital and smart cities are often characterized by arising M2M applications. However, a city is, before all, the gathering of citizens, who use digital services and mobile Internet for increasing their quality of life, empowerment, and entertainment opportunities. Some data flows should be gathered to, or distributed from, an information system. Some other should be disseminated to a geographically or time constrained perimeter. Future usage may induce peer-to-peer like traffics. Moreover these services are also an enabler of new usages of the urban environment. Solutions built within the capillary network paradigm have to manage this heterogeneity of traffic requirements and user behaviors.

By following these guidelines, the UrbaNet ambition is to go one step beyond traditional approaches discussed above. The capillary network paradigm for Smart Cities is tightly linked to the specificities of the metropolitan context and the citizens' activity. Our proposal is thus to re-think the way capillary network technologies are developed, considering a broader and more practical perspective.

3.3. Characterizing urban networks

Our first objective is to understand and model those properties of real-world urban environments that have an impact on the design, deployment and operation of capillary networks. It means to collect and analyze data from actual deployments and services, as well as testbeds experiments. These data have then to be correlated with urban characteristics, e.g. topography, density of population and activities. The objective is to deduce analytical models, simulations and traces of realistic scenarios that can be leveraged afterward. We structure the axis into three tasks that correspond to the three broad categories of networking aspects affected by the urban context.

- **Topological characteristics**. Nowadays, the way urban wireless network infrastructures are typically represented in the literature is dissatisfying. As an example, wireless links are mostly represented as symmetric, lossless channels whose signal quality depends continuously on the distance between the transmitter and the receiver. No need to say, real-world behaviors are very far from

these simplified representations. Another example, topologies are generally modeled according to deterministic (e.g., regular grids and lattices, or perfect hexagonal cell coverages) or stochastic (e.g., random uniform distributions over unbound surfaces) approaches. These make network problems mathematically tractable and simulations easier to set up, but are hardly representative of the layouts encountered in the real world. Employing simplistic models helps understanding some fundamental principles but risks to lead to unreliable results, both from the viewpoint of the network architecture design and from that of its performance evaluation. It is thus our speculation that the actual operations and the real-world topologies of infrastructured capillary networks are key to the successful deployment of these technologies, and, in this task, we aim at characterizing them. To that end, we leverage existing collaborations with device manufacturers (Alcatel-Lucent, HiKob) and operators (Orange), as well as collaboration such as the Sense City project and testbed experiments, in order to provide models that faithfully mimic the behavior of real world network devices. The goal is to understand the important features of the topologies, including, e.g., their overall connectivity level, spatial density, degree distribution, regularity, etc. Building on these results, we try to define network graph models that reproduce such major features and can be employed for the development and evaluation of capillary network solutions.

- **Mobilities**. We aim at understanding and modeling the mobile portion of capillary networks as well as the impact of the human mobility on the network usage. Our definition of "mobile portion" includes traditional mobile users as well as all communication-enabled devices that autonomously interact with Internet-based servers and among themselves. There have been efforts to collect real-world movement traces, to generate synthetic mobility dataset and to derive mobility models. However, real-world traces remain limited to small scenarios or circumstantial subsets of the users (e.g., cabs instead of the whole road traffic). Synthetic traces are instead limited by their scale and by their level of realism, still insufficient. Finally, even the most advanced models cannot but provide a rough representation of user mobility in urban areas, as they do not consider the street layout or the human activity patterns. In the end, although often deprecated, random or stochastic mobility models (e.g., random walks, exponential inter-arrivals and cell residence times) are still the common practice. We are well aware of the paramount importance of a faithful representation of device and user mobility within capillary networks and, in order to achieve it, we leverage a number of realistic sources, including Call Detail Records (CDR) collected by mobile operators, Open Data initiatives, real-world social network data, and experiments. We collect data and analyze it, so as to infer the critical properties of the underlying mobility patterns.
- **Data traffic patterns**. The characterization of capillary network usages means understanding and modeling when, where and how the wireless access provided by the diverse capillary network technologies is exploited by users and devices. In other words, we are interested in learning which applications are used at different geographical locations and day times, which urban phenomena generate network usage, and which kind of data traffic load they induce on the capillary network. Properly characterizing network usages is as critical as correctly modeling network topology and mobility. Indeed, the capillary networks being the link directly collecting the data from end devices, we cannot count on statistical smoothing which yields regular distributions. Unfortunately, the common practice is to consider, e.g., that each user or device generates a constant data traffic or follows on/off models, that the offered load is uniform over space and does not vary over time, that there is small difference between uplink and downlink behaviors, or that source/destination node pairs are randomly distributed in the network. We plan to go further on the specific scenarios we address, such as smart-parking, floating car data, tele-metering, road traffic management of pollution detection. To that end, we collect real-world data, explore it and derive properties useful to the accurate modeling of content consumption.

3.4. Autonomic networking protocols

While the capillary networks concept covers a large panel of technologies, network architectures, applications and services, common challenges remain, regardless the particular choice of a technology or architecture.

Our record of research on spontaneous and multi-hop networks let us think that autonomic networking appears as the main issue: the connectivity to Internet, to cyber-physical systems, to Information Systems should be transparent for the user, context-aware and location-aware. To address these challenges, a capillary network model is required. Unfortunately, very few specific models fit this task today. However, a number of important, specific capillary networks properties can already be inferred from recent experiments: distributed and localized topologies, very high node degree, dynamic network diameter, unstable / asymmetric / non-transitive radio links, concurrent topologies, heterogeneous capabilities, etc. These properties can already be acknowledged in the design of networking solutions, and they are particularly challenging for the functioning of the MAC layer and QoS support. Clearly, capillary networks provide new research opportunities with regard to networking protocols design.

- **Self-* protocols**. In this regard, self-configuration, self-organization and self-healing are some of the major concerns within the context of capillary networks. Solving such issues would allow spontaneous topologies to appear dynamically in order to provide a service depending of the location and the context, while also adapting to the interactions imposed by the urban environment. Moreover, these mechanisms have the capacity to alleviate the management of the network and the deployment engineering rules, and can provide efficient support to the network dynamics due to user mobility, environment modifications, etc. The designed protocols have to be able to react to traffic requests and local node densities. We address such self-adaptive protocols as a transversal solution to several scenarios, e.g. pollution monitoring, smart-services depending on human activities, vehicle to infrastructure communications, etc. In architectures where self-* mechanisms govern the protocol design, both robustness and energy are more than ever essential challenges at the network layer. Solutions such as energy-harvesting can significantly increase the network lifetime in this case, therefore we investigate their impact on the mechanisms at both MAC and network layers.
- **Quality of service issues**. The capillary networks paradigm implies a simultaneous deployment of multiple wireless technologies, and by different entities (industry, local community, citizens). This means that some applications and services can be provided concurrently by different parts of the capillary network, while others might require the cooperation of multiple parties. The notion of Service Level Agreement (SLA) for traffic differentiation, quality of service support (delay, reliability, etc.) is a requirement in these cases for scalability purposes and resource sharing. We contribute to a proper definition of this notion and the related network mechanisms in the settings of low power wireless devices. Because of the urban context, but also because of the wireless media itself, network connectivity is always temporary, while applications require a delivery ratio close to 100%. We investigate different techniques that can achieve this objective in an urban environment.
- **Data impact**. Capillary networks suffer from low capacity facing the increasing user request. In order to cope with network saturation, a promising strategy is to consider the nature of the transmitted data in the development of the protocols. Data aggregation and data gathering are two concepts with a major role to play in this context of limited capacity. In particular, combining local aggregation and measurement redundancy for improving on data reliability is a promising idea, which can also be important for energy saving purposes. Even if the data flow is well known and regular, e.g. temperature or humidity metering, developing aggregation schemes tailored to the constraints of the urban environment is a challenge we address within the UrbaNet team. Many urban applications generate data which has limited spatial and temporal perimeters of relevance, e.g. smart-parking applications, community information broadcasting, etc. When solely a spatial range of relevance is considered, the underlying mechanisms are denoted “geocasting”. We also address these spatio-temporal constraints, which combine geocasting approaches with real-time techniques.

3.5. Optimizing cellular network usage

The capacity of cellular networks, even those that are now being planned, does not seem able to cope with the increasing demands of data users. Moreover, new applications with high bandwidth requirements are also foreseen, for example in the intelligent transportation area, and an exponential growth in signaling traffic is

expected in order to enable this data growth. Cumulated with the lack of available new spectrum, this leads to an important challenge for mobile operators, who are looking at both licensed and unlicensed technologies for solutions. The usual strategy consists in a dramatic densification of micro-cells coverage, allowing both to minimize the transmission power of cellular networks as well as to increase the network capacity. However, this solution has obvious physical limits, which we work on determining, and we propose exploiting the capillarity of network interfaces as a complementary solution.

- **Green cellular network**. Increasing the density of micro-cells means multiplying the energy consumption issues. Indeed, the energy consumption of actual LTE eNodeBs and relays, whatever their state, idle, transmitting or receiving, is a major and growing part of the access network energy consumption. For a sustainable deployment of such micro-cell infrastructures and for a significative decrease of the overall energy consumption, an operator needs to be able to switch off cells when they are not absolutely needed. The densification of the cells induces the need for an autonomic control of the on/off state of cells. One solution in this sense can be to adapt the WSN mechanisms to the energy models of micro-cells and to the requirements of a cellular network. The main difficulty here is to be able to adapt and assess the proposed solutions in a realistic environment (in terms of radio propagation, deployment of the cells, user mobility and traffic dynamics).
- **Offloading**. Offloading the cellular infrastructure implies taking advantage of the wealth of connectivity provided by capillary networks instead of relying solely on 4G connectivity. Cellular operators usually possess an important ADSL or cable infrastructure for wired services, the development of femtocell solutions thus becomes very popular. However, while femtocells can be an excellent solution in zones with poor coverage, their extensive use in areas with a high density of mobile users leads to serious interference problems that are yet to be solved. Taking advantage of capillarity for offloading cellular data relies on using IEEE 802.11 Wi-Fi (or other similar technologies) access points or direct device-to-device communications. The ubiquity of Wi-Fi access in urban areas makes this solution particularly interesting, and many studies have focused on its potential. However, these studies fail to take into account the usually low quality of Wi-Fi connections in public areas, and they consider that a certain data rate can be sustained by the Wi-Fi network regardless of the number of contending nodes. In reality, most public Wi-Fi networks are optimized for connectivity, but not for capacity, and more research in this area is needed to correctly assess the potential of this technology. Direct opportunistic communication between mobile users can also be used to offload an important amount of data. This solution raises a number of major problems related to the role of social information and multi-hop communication in the achievable offload capacity. Moreover, in this case the business model is not yet clear, as operators would indeed offload traffic, but also lose revenue as direct ad-hoc communication would be difficult to charge and privacy issues may arise. However, combining hotspot connectivity and multi-hop communications is an appealing answer to broadcasting geo-localized informations efficiently.

WHISPER Team

3. Research Program

3.1. Scientific Foundations

3.1.1. Program analysis

A fundamental goal of the research in the Whisper team is to elicit and exploit the knowledge found in existing code. To do this in a way that scales to a large code base, systematic methods are needed to infer code properties. We may build on either static [38], [40], [41] or dynamic analysis [57], [59], [64]. Static analysis consists of approximating the behavior of the source code from the source code alone, while dynamic analysis draws conclusions from observations of sample executions, typically of test cases. While dynamic analysis can be more accurate, because it has access to information about actual program behavior, obtaining adequate test cases is difficult. This difficulty is compounded for infrastructure software, where many, often obscure, cases must be handled, and external effects such as timing can have a significant impact. Thus, we expect to primarily use static analyses. Static analyses come in a range of flavors, varying in the extent to which the analysis is *sound*, *i.e.*, the extent to which the results are guaranteed to reflect possible run-time behaviors.

One form of sound static analysis is *abstract interpretation* [40]. In abstract interpretation, atomic terms are interpreted as sound abstractions of their values, and operators are interpreted as functions that soundly manipulate these abstract values. The analysis is then performed by interpreting the program in a compositional manner using these abstracted values and operators. Alternatively, *dataflow analysis* [48] iteratively infers connections between variable definitions and uses, in terms of local transition rules that describe how various kinds of program constructs may impact variable values. Schmidt has explored the relationship between abstract interpretation and dataflow analysis [69]. More recently, more general forms of symbolic execution [38] have emerged as a means of understanding complex code. In symbolic execution, concrete values are used when available, and these are complemented by constraints that are inferred from terms for which only partial information is available. Reasoning about these constraints is then used to prune infeasible paths, and obtain more precise results. A number of works apply symbolic execution to operating systems code [35], [36].

While sound approaches are guaranteed to give correct results, they typically do not scale to the very diverse code bases that are prevalent in infrastructure software. An important insight of Engler et al. [43] was that valuable information could be obtained even when sacrificing soundness, and that sacrificing soundness could make it possible to treat software at the scales of the kernels of the Linux or BSD operating systems. Indeed, for certain types of problems, on certain code bases, that may mostly follow certain coding conventions, it may mostly be safe to *e.g.*, ignore the effects of aliases, assume that variable values are unchanged by calls to unanalyzed functions, etc. Real code has to be understood by developers and thus cannot be too complicated, so such simplifying assumptions are likely to hold in practice. Nevertheless, approaches that sacrifice soundness also require the user to manually validate the results. Still, it is likely to be much more efficient for the user to perform a potentially complex manual analysis in a specific case, rather than to implement all possible required analyses and apply them everywhere in the code base. A refinement of unsound analysis is the CEGAR approach [39], in which a highly approximate analysis is complemented by a sound analysis that checks the individual reports of the approximate analysis, and then any errors in reasoning detected by the sound analysis are used to refine the approximate analysis. The CEGAR approach has been applied effectively on device driver code in tools developed at Microsoft [27]. The environment in which the driver executes, however, is still represented by possibly unsound approximations.

Going further in the direction of sacrificing soundness for scalability, the software engineering community has recently explored a number of approaches to code understanding based on techniques developed in the areas of natural language understanding, data mining, and information retrieval. These approaches view code, as well as other software-reated artifacts, such as documentation and postings on mailing lists, as bags of words structured in various ways. Statistical methods are then used to collect words or phrases that seem to be highly correlated, independently of the semantics of the program constructs that connect them. The obliviousness to program semantics can lead to many false positives (invalid conclusions) [53], but can also highlight trends that are not apparent at the low level of individual program statements. We have explored combining such statistical methods with more traditional static analysis in identifying faults in the usage of constants in Linux kernel code [52].

3.1.2. Domain Specific Languages

Writing low-level infrastructure code is tedious and difficult, and verifying it is even more so. To produce non-trivial programs, we could benefit from moving up the abstraction stack for both programming and proving as quickly as possible. Domain-specific languages (DSLs), also known as *little languages*, are a means to that end [5] [61].

3.1.2.1. Traditional approach.

Using little languages to aid in software development is a tried-and-trusted technique [71] by which programmers can express high-level ideas about the system at hand and avoid writing large quantities of formulaic C boilerplate.

This approach is typified by the Devil language for hardware access [7]. An OS programmer describes the register set of a hardware device in the high-level Devil language, which is then compiled into a library providing C functions to read and write values from the device registers. In doing so, Devil frees the programmer from having to write extensive bit-manipulation macros or inline functions to map between the values the OS code deals with, and the bit-representation used by the hardware: Devil generates code to do this automatically.

However, DSLs are not restricted to being "stub" compilers from declarative specifications. The Bossa language [6] is a prime example of a DSL involving imperative code (syntactically close to C) while offering a high-level of abstraction. This design of Bossa enables the developer to implement new process scheduling policies at a level of abstraction tailored to the application domain.

Conceptually, a DSL both abstracts away low-level details and justifies the abstraction by its semantics. In principle, it reduces development time by allowing the programmer to focus on high-level abstractions. The programmer needs to write less code, in a language with syntax and type checks adapted to the problem at hand, thus reducing the likelihood of errors.

3.1.2.2. Embedding DSLs.

The idea of a DSL has yet to realize its full potential in the OS community. Indeed, with the notable exception of interface definition languages for remote procedure call (RPC) stubs, most OS code is still written in a low-level language, such as C. Where DSL code generators are used in an OS, they tend to be extremely simple in both syntax and semantics. We conjecture that the effort to implement a given DSL usually outweighs its benefit. We identify several serious obstacles to using DSLs to build a modern OS: specifying what the generated code will look like, evolving the DSL over time, debugging generated code, implementing a bug-free code generator, and testing the DSL compiler.

Filet-o-Fish (FoF) [3] addresses these issues by providing a framework in which to build correct code generators from semantic specifications. This framework is presented as a Haskell library, enabling DSL writers to *embed* their languages within Haskell. DSL compilers built using FoF are quick to write, simple, and compact, but encode rigorous semantics for the generated code. They allow formal proofs of the run-time behavior of generated code, and automated testing of the code generator based on randomized inputs, providing greater test coverage than is usually feasible in a DSL. The use of FoF results in DSL compilers that OS developers can quickly implement and evolve, and that generate provably correct code. FoF has been used

to build a number of domain-specific languages used in Barrelfish, [29] an OS for heterogeneous multicore systems developed at ETH Zurich.

The development of an embedded DSL requires a few supporting abstractions in the host programming language. FoF was developed in the purely functional language Haskell, thus benefiting from the type class mechanism for overloading, a flexible parser offering convenient syntactic sugar, and purity enabling a more algebraic approach based on small, composable combinators. Object-oriented languages – such as Smalltalk [44] and its descendant Pharo [32] – or multi-paradigm languages – such as the Scala programming language [63] – also offer a wide range of mechanisms enabling the development of embedded DSLs. Perhaps suprisingly, a low-level imperative language – such as C – can also be extended so as to enable the development of embedded compilers [30].

3.1.2.3. Certifying DSLs.

Whilst automated and interactive software verification tools are progressively being applied to larger and larger programs, we have not yet reached the point where large-scale, legacy software – such as the Linux kernel – could formally be proved "correct". DSLs enable a pragmatic approach, by which one could realistically strengthen a large legacy software by first narrowing down its critical component(s) and then focus our verification efforts onto these components.

Dependently-typed languages, such as Coq or Idris, offer an ideal environment for embedding DSLs [37], [33] in a unified framework enabling verification. Dependent types support the type-safe embedding of object languages and Coq's mixfix notation system enables reasonably idiomatic domain-specific concrete syntax. Coq's powerful abstraction facilities provide a flexible framework in which to not only implement and verify a range of domain-specific compilers [3], but also to combine them, and reason about their combination.

Working with many DSLs optimizes the "horizontal" compositionality of systems, and favors reuse of building blocks, by contrast with the "vertical" composition of the traditional compiler pipeline, involving a stack of comparatively large intermediate languages that are harder to reuse the higher one goes. The idea of building compilers from reusable building blocks is a common one, of course. But the interface contracts of such blocks tend to be complex, so combinations are hard to get right. We believe that being able to write and verify formal specifications for the pieces will make it possible to know when components can be combined, and should help in designing good interfaces.

Furthermore, the fact that Coq is also a system for formalizing mathematics enables one to establish a close, formal connection between embedded DSLs and non-trivial domain-specific models. The possibility of developing software in a truly "model-driven" way is an exciting one. Following this methodology, we have implemented a certified compiler from regular expressions to x86 machine code [4]. Interestingly, our development crucially relied on an existing Coq formalization, due to Braibant and Pous, [34] of the theory of Kleene algebras.

While these individual experiments seem to converge toward embedding domain-specific languages in rich type theories, further experimental validation is required. Indeed, Barrelfish is an extremely small software compared to the Linux kernel. The challenge lies in scaling this methodology up to large software systems. Doing so calls for a unified platform enabling the development of a myriad of DSLs, supporting code reuse across DSLs as well as providing support for mechanically-verified proofs.

3.2. Research direction: developing drivers using Genes

We believe that weaknesses of previous methods for easing device driver development arise from an insufficient understanding of the range and scope of driver functionality, as required by real devices and OSes. We propose a new methodology for understanding device drivers, inspired by the biological field of genomics. Rather than focusing on the input/output behavior of a device, we take the radically new methodology of studying existing device driver code itself. On the one hand, this methodology makes it possible to identify the behaviors performed by real device drivers, whether to support the features of the device and the OS, or to improve properties such as safety or performance. On the other hand, this methodology makes it possible to capture the actual patterns of code used to implement these behaviors, raising the level of abstraction from

individual operations to collections of operations implementing a single functionality, which we refer to as *genes*. Because the requirements of the device remain fixed, regardless of the OS, we expect to find genes with common behaviors across different OSes, even when those genes have a different internal structure. This leads to a view of a device driver as being constructed as a composition of genes, thus opening the door to new methodologies to address the problems faced by real driver developers. Among these, we have so far identified the problems of developing drivers, porting existing drivers to other OSes, backporting existing drivers to older OS versions, and long-term maintenance of the driver code.

Our short term goal is to "sequence" the complete set of genes for a set of related drivers. In the longer term, we plan to develop methodologies based on genes for aiding in driver development and maintenance. This work is currently financed by a grant from the Direction Générale de l'Armement (DGA) that supports the PhD of Peter Senna Tschudin. Valentin Rothberg's PhD is supported by an Inria Cordi-S grant.

3.3. Research direction: developing infrastructure software using Domain Specific Languages

We wish to pursue a *declarative* approach to developing infrastructure software. Indeed, there exists a significant gap between the high-level objectives of these systems and their implementation in low-level, imperative programming languages. To bridge that gap, we propose an approach based on domain-specific languages (DSLs). By abstracting away boilerplate code, DSLs increase the productivity of systems programmers. By providing a more declarative language, DSLs reduce the complexity of code, thus the likelihood of bugs.

Traditionally, systems are built by accretion of several, independent DSLs. For example, one might use Devil [7] to interact with devices, Bossa [6] to implement the scheduling policies, and Zebu [2] to implement some networking protocols. However, much effort is duplicated in implementing the back-ends of the individual DSLs. Our long term goal is to design a unified framework for developing and composing DSLs, following our work on Filet-o-Fish [3]. By providing a single conceptual framework, we hope to amortize the development cost of a myriad of DSLs through a principled approach to reusing and composing DSLs.

Beyond the software engineering aspects, a unified platform brings us closer to the implementation of mechanically-verified DSLs. Dagand's recent work using the Coq proof assistant as an x86 macro-assembler [4] is a step in that direction, which belongs to a larger trend of hosting DSLs in dependent type theories [33], [62], [37]. A key benefit of those approaches is to provide – by construction – a formal, mechanized semantics to the DSLs thus developed. This semantics offers a foundation on which to base further verification efforts, whilst allowing interaction with non-verified code. We advocate a methodology based on incremental, piece-wise verification. Whilst building fully-certified systems from the top-down is a worthwhile endeavor [49], we wish to explore a bottom-up approach by which one focuses first and foremost on crucial subsystems and their associated properties.

We plan to apply this methodology for implementing a certified DSL for describing serializers and deserializers of binary datastreams. This work will build on our experience in designing Zebu [2], a DSL for describing text-based protocols. Inspired by our experience implementing a certified regular expression compiler in x86 [4], we wish to extend Zebu to manipulate binary data. Such a DSL should require a single description of a binary format and automatically generate a serializer/deserializer pair. This dual approach – relating a binary format to its semantic model – is inspired by the Parsifal [54] and Nail [28] format languages. A second challenge consists in guaranteeing the functional correctness of the serializer/deserializer pair generated by the DSL: one would wish to prove that any serialized data can be deserialized to itself, and conversely. The RockSalt's project [62] provides the conceptual tools, in a somewhat simpler setting, to address this question.

Packet filtering is another sweet spot for DSLs. First, one needs a DSL for specifying the filtering rules. This is standard practice [60]. However, in our attempt to establish the correctness of the packet filter, we will be led to equip this DSL with a mechanized semantics, formally describing the precise meaning of each construct of the language. Second, packet filters are usually implemented through a matching engine that is, essentially, a bytecode interpreter. To establish the correctness of the packet filter, we shall then develop a mechanized semantics of this bytecode and prove that the *compilation* from filtering rules to bytecode

preserves the intended semantics. Because a packet filter lies at the entry-point of a network, safety is crucial: we would like to guarantee that the packet filter cannot crash and is not vulnerable to an attack. Beyond mere safety, functional correctness is essential too: we must guarantee that the high-level filtering rules are indeed applied as expected by the matching engine. A loophole in the compilation could leave the network open to an attack or prevent legitimate traffic from reaching its destination. Finally, the safety of the packet filter *cannot* be established at the expense of performance. Indeed, if the packet filter were to become a bottleneck, the infrastructure it aimed at protecting would easily become subject to Denial of Service (DoS) attacks. Filtering rules should therefore be compiled efficiently: the corresponding optimizations will have to be verified [73].

ALICE Project-Team

3. Research Program

3.1. Introduction

Computer Graphics is a quickly evolving domain of research. These last few years, both acquisition techniques (e.g., range laser scanners) and computer graphics hardware (the so-called GPU's, for Graphics Processing Units) have made considerable advances. However, despite these advances, fundamental problems still remain open. For instance, a scanned mesh composed of hundred million triangles cannot be used directly in real-time visualization or complex numerical simulation. To design efficient solutions for these difficult problems, ALICE studies two fundamental issues in Computer Graphics:

- the representation of the objects, i.e., their geometry and physical properties;
- the interaction between these objects and light.

Historically, these two issues have been studied by independent research communities. However, we think that they share a common theoretical basis. For instance, multi-resolution and wavelets were mathematical tools used by both communities [28]. We develop a new approach, which consists in studying the geometry and lighting from the *numerical analysis* point of view. In our approach, geometry processing and light simulation are systematically restated as a (possibly non-linear and/or constrained) functional optimization problem. This type of formulation leads to algorithms that are more efficient. Our long-term research goal is to find a formulation that permits a unified treatment of geometry and illumination over this geometry.

3.2. Geometry Processing for Engineering

Keywords: Mesh processing, parameterization, splines

Geometry processing recently emerged (in the middle of the 90's) as a promising strategy to solve the geometric modeling problems encountered when manipulating meshes composed of hundred millions of elements. Since a mesh may be considered to be a *sampling* of a surface - in other words a *signal* - the *digital signal processing* formalism was a natural theoretic background for this subdomain (see e.g., [29]). Researchers of this domain then studied different aspects of this formalism applied to geometric modeling.

Although many advances have been made in the geometry processing area, important problems still remain open. Even if shape acquisition and filtering is much easier than 30 years ago, a scanned mesh composed of hundred million triangles cannot be used directly in real-time visualization or complex numerical simulation. For this reason, automatic methods to convert those large meshes into higher level representations are necessary. However, these automatic methods do not exist yet. For instance, the pioneer Henri Gouraud often mentions in his talks that the *data acquisition* problem is still open. Malcolm Sabin, another pioneer of the "Computer Aided Geometric Design" and "Subdivision" approaches, mentioned during several conferences of the domain that constructing the optimum control-mesh of a subdivision surface so as to approximate a given surface is still an open problem. More generally, converting a mesh model into a higher level representation, consisting of a set of equations, is a difficult problem for which no satisfying solutions have been proposed. This is one of the long-term goals of international initiatives, such as the AIMShape European network of excellence.

Motivated by gridding application for finite elements modeling for oil and gas exploration, in the frame of the Gocad project, we started studying geometry processing in the late 90's and contributed to this area at the early stages of its development. We developed the LSCM method (Least Squares Conformal Maps) in cooperation with Alias Wavefront [24]. This method has become the de-facto standard in automatic unwrapping, and was adopted by several 3D modeling packages (including Maya and Blender). We experimented various applications of the method, including normal mapping, mesh completion and light simulation [2].

However, classical mesh parameterization requires to partition the considered object into a set of topological disks. For this reason, we designed a new method (Periodic Global Parameterization) that generates a continuous set of coordinates over the object [5]. We also showed the applicability of this method, by proposing the first algorithm that converts a scanned mesh into a Spline surface automatically [4].

We are still not fully satisfied with these results, since the method remains quite complicated. We think that a deeper understanding of the underlying theory is likely to lead to both efficient and simple methods. For this reason, in 2012 we studied several ways of discretizing partial differential equations on meshes, including Finite Element Modeling and Discrete Exterior Calculus. In 2013, we also explored Spectral Geometry Processing and Sampling Theory (more on this below).

3.3. Computer Graphics

Keywords: texture synthesis, shape synthesis, texture mapping, visibility

Content creation is one of the major challenges in Computer Graphics. Modeling shapes and surface appearances which are visually appealing and at the same time enforce precise design constraints is a task only accessible to highly skilled and trained designers.

In this context the team focuses on methods for by-example content creation. Given an input example and a set of constraints, we design algorithms that can automatically generate a new shape (geometry+texture). We formulate the problem of content synthesis as the joint optimization of several objectives: Preserving the local appearance of the example, enforcing global objectives (size, symmetries, mechanical properties), reaching user defined constraints (locally specified geometry, contacts). This results in a wide range of optimization problems, from statistical approaches (Markov Random fields), to combinatorial and linear optimization techniques.

As as complement to the design of techniques for automatic content creation, we also work on the representation of the content, so as to allow for its efficient manipulation. In this context we develop data-structures and algorithms targeted at massively parallel architectures, such as GPUs. These are critical to reach the interactive rates expected from a content creation technique. We also propose novel ways to store and access content stored along surfaces [6] or in volumes [1] [23].

The team also continues research in core topics of computer graphics at the heart of realistic rendering and realistic light simulation techniques; for example, mapping textures on surfaces, or devising visibility relationships between 3D objects populating space.

ALPAGE Project-Team

3. Research Program

3.1. From programming languages to linguistic grammars

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Pierre Boullier, Djamé Seddah, Corentin Ribeyre.

Historically, several members of Alpage were originally specialists in the domain of modeling and parsing for programming languages, and have been working for more than 15 years on the generalization and extension of the techniques involved to the domain of natural language. The shift from programming language grammars to NLP grammars seriously increases complexity (e.g., grammar size [0]) and requires ways to handle the ambiguities inherent in every human language. It is well known that these ambiguities are the sources of many badly handled combinatorial explosions.

Furthermore, while most programming languages are expressed by (subclasses) of well-understood context-free grammars (CFGs), no consensual grammatical formalism has yet been accepted by the whole linguistic community for the description of human languages. On the contrary, new formalisms (or variants of older ones) appear constantly. Many of them may be classified into the three following large families:

Mildly Context-Sensitive (MCS) formalisms They manipulate possibly complex elementary structures with enough restrictions to ensure the possibility of parsing with polynomial time complexities. They include, for instance, Tree Adjoining Grammars (TAGs) and Multi-component TAGs with trees as elementary structures, Linear Indexed Grammars (LIGs). Although they are strictly more powerful than MCS formalisms, Range Concatenation Grammars (RCGs, introduced and used by Alpage members, such as Pierre Boullier and Benoît Sagot [59], [108], [116]) are also parsable in polynomial time.

Unification-based formalisms They combine a context-free backbone with logic arguments as decoration on non-terminals. Most famous representatives are Definite Clause Grammars (DCGs) where PROLOG powerful unification is used to compute and propagate these logic arguments. More recent formalisms, like Lexical Functional Grammars (LFGs) and Head-Driven Phrasal Structure Grammars (HPSGs) rely on more expressive Typed Feature Structures (TFS) or constraints.

Unification-based formalisms with an MCS backbone The two above-mentioned characteristics may be combined, for instance by adding logic arguments or constraints to non-terminals in TAGs.

An efficient way to develop large-coverage hand-crafted symbolic grammars is to use adequate tools and adequate levels of representation, and in particular Meta-Grammars, one of Alpage's areas of expertise, especially with the FRMG grammar and parser for French based on the DyALog logic programming environment [136], [130]. Meta-Grammars (MGs) allows the linguist to focus on a modular description of the linguistic aspects of a grammar, rather than focusing on the specific aspects of a given grammatical formalism. Translation from MGs to grammatical formalisms such as TAG or LFG may be automatically handled. Graphical environments can be used to design MGs and their modularity provides a promising way for sharing the description of common linguistic phenomena across human languages.

3.2. Statistical Parsing

Participants: Djamé Seddah, Marie-Hélène Candito, Benoit Crabbé, Éric Villemonte de La Clergerie, Benoît Sagot, Corentin Ribeyre, Pierre Boullier, Maximin Coavoux.

[0] boullier:2010:inria-00516341:1

Contrary to symbolic approaches to parsing, in statistical parsing, the grammar is extracted from a corpus of syntactic trees : a treebank. The main advantage of the statistical approach is to encode within the same framework the parsing and disambiguating tasks. The extracted grammar rules are associated with probabilities that allow to score and rank the output parse trees of an input sentence. This obvious advantage of probabilistic context-free grammars has long been counterbalanced by two main shortcomings that resulted in poor performance for plain PCFG parsers: (i) the generalization encoded in non terminal symbols that stand for syntagmatic phrases is too coarse (so probabilistic independence between rules is too strong an assertion) and (ii) lexical items are underused. In the last decade though, effective solutions to these shortcomings have been proposed. Symbol annotation, either manual [87] or automatic [100], [101] captures inter-dependence between CFG rules. Lexical information is integrated in frameworks such as head-driven models that allow lexical heads to percolate up the syntagmatic tree [73], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [71].

In the same period, totally different parsing architectures have been proposed, to obtain dependency-based syntactic representations. The properties of dependency structures, in which each word is related to exactly one other word, make it possible to define dependency parsing as a sequence of simple actions (such as read buffer and store word on top of a stack, attach read word as dependent of stack top word, attach read word as governor of stack top word ...) [140], [98]. Classifiers can be trained to choose the best action to perform given a partial parsing configuration. In another approach, dependency parsing is cast into the problem of finding the maximum spanning tree within the graph of all possible word-to-word dependencies, and online classification is used to weight the edges [92]. These two kinds of statistical dependency parsing allow to benefit from discriminative learning, and its ability to easily integrate various kinds of features, which is typically needed in a complex task such as parsing.

Statistical parsing is now effective, both for syntagmatic representations and dependency-based syntactic representations. Alpage has obtained state-of-the-art parsing results for French, by adapting various parser learners for French, and works on the current challenges in statistical parsing, namely (1) robustness and portability across domains and (2) the ability to incorporate exogenous data to improve parsing attachment decisions. Alpage is the first French team to have turned the French TreeBank into a resource usable for training statistical parsers, to distribute a dependency version of this treebank, and to make freely available various state-of-the art statistical POS-taggers and parsers for French. We review below the approaches that Alpage has tested and adapted, and the techniques that we plan to investigate to answer these challenges.

In order to investigate statistical parsers for French, we have first worked how to use the French Treebank [55], [54] and derive the best input for syntagmatic statistical parsing [75]. Benchmarking several PCFG-based learning frameworks [122] has led to state-of-the-art results for French, the best performance being obtained with the split-merge Berkeley parser (PCFG with latent annotations) [101].

In parallel to the work on dependency based representation, presented in the next paragraph, we also conducted a preliminary set of experiments on richer parsing models based on Stochastic Tree Insertion Grammars as used in [71] and which, besides their inferior performance compared to PCFG-LA based parser, raise promising results with respect to dependencies that can be extracted from derivation trees. One variation we explored, that uses a specific TIG grammar instance, a *vertical* grammar called *spinal* grammars, exhibits interesting properties wrt the grammar size typically extracted from treebanks (a few hundred unlexicalized trees, compared to 14 000 CFG rules). These models are currently being investigated in our team [126].

Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [67], that are easier to use for various NLP applications such as question-answering or information extraction, and that are a better ground for further semantic analysis. This conversion can be applied on the treebank, before training a dependency-based parser, or on PCFG-LA parsed trees. This gives the possibility to evaluate and compare on the same gold data, both syntagmatic- and dependency-based statistical parsing. This also paved the way for studies on the influence of various types of lexical information.

3.3. Robust linguistic processing

Participants: Djamé Seddah, Benoît Sagot, Éric Villemonte de La Clergerie, Marie-Hélène Candito, Kata Gábor, Pierre Magistry, Marion Baranes.

The constitution of resources such as lexica or grammars raises the issues of the evaluation of these resources to assess their quality and coverage. For this reason, Alpage was the leader of the PASSAGE ANR project (ended in June 2010), which is the follow-up of the EASy parsing evaluation campaign held in 2004 and conducted by team LIR at LIMSI.

However, although developing parsing techniques, grammars (symbolic or probabilistic), and lexica constitute the key efforts towards deep large-scale linguistic processing, these components need to be included inside a full and robust processing chain, able to handle any text from any source, especially out-of-domain text genres. Such texts that exhibit properties (e.g., lexical and syntactic properties) that are different or differently distributed than what is found on standard data (e.g., training corpora for statistical parsers). The development of shallow processing chains, such as SxPipe (see 5.5), is not a trivial task [110]. Obviously, they are often used as such, and not only as pre-processing tools before parsing, since they perform the basic tasks that produce immediately usable results for many applications, such as tokenization, sentence segmentation, spelling correction (e.g., for improving the output of OCR systems), named entity detection, disambiguation and resolution, as well as morphosyntactic tagging.

Still, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains. This is especially the case, beyond the standard out-of-domain corpora mentioned above, for user-generated content. Indeed, until very recently out-of-domain text genres that have been prioritized have not been Web 2.0 sources, but rather biomedical texts, child language and general fiction (Brown corpus). Adaptation to user-generated content is a particularly difficult instance of the domain adaptation problem since Web 2.0 is not really a domain: it consists of utterances that are often ungrammatical. It even shares some similarities with spoken language [129]. The poor overall quality of texts found on such media lead to weak parsing and even POS-tagging results. This is because user-generated content exhibits both the same issues as other out-of-domain data, but also tremendous issues related to tokenization, typographic and spelling issues that go far beyond what statistical tools can learn from standard corpora. Even lexical specificities are often more challenging than on edited out-of-domain text, as neologisms built using productive morphological derivation, for example, are less frequent, contrarily to slang, abbreviations or technical jargon that are harder to analyse and interpret automatically.

In order to fully prepare a shift toward more robustness, we developed a first version of a richly annotated corpus of user-generated French text, the French Social Media Bank [7], which includes not only POS, constituency and functional information, but also a layer of "normalized" text. This corpus is fully available and constitutes the first data set on Facebook data to date and the first instance of user generated content for a morphologically-rich language. Thanks to the support of the Labex EFL through, we are currently the finalizing the second release of this data set, extending toward a full treebank of over 4,000 sentences (see section 6.9).

Besides delivering a new data set, our main purpose here is to be able to compare two different approaches to user-generated content processing: either training statistical models on the original annotated text, and use them on raw new text; or developing normalization tools that help improving the consistency of the annotations, train statistical models on the normalized annotated text, and use them on normalized texts (before un-normalizing them).

However, this raises issues concerning the normalization step. A good sandbox for working on this challenging task is that of POS-tagging. For this purpose, we did leverage Alpage's work on MElt, a state-of-the art POS tagging system [80] (see 5.5). A first round of experiments on English have already led to promising results during the shared task on parsing user-generated content organized by Google in May 2012 [102], as Alpage was ranked second and third [125]. For achieving this result, we brought together a preliminary implementation of a normalization wrapper around the MElt POS tagger followed by a state-of-the art statistical parser improved by several domain adaptation techniques we originally developed for parsing edited out-of-domain texts. Those techniques are based on the unsupervized learning of word clusters *a la* Brown and benefit

from morphological treatments (such as lemmatization or desinflection) [123]. More recent developments are sketched in section 4.2

One of our objectives is to generalize the use of the normalization wrapper approach to both POS tagging and parsing, for English and French, in order to improve the quality of the output parses. However, this raises several challenges: non-standard contractions and compounds lead to unexpected syntactic structures. A first round of experiments on the French Social Media Bank showed that parsing performance on such data are much lower than expected. This is why, we are actively working to improve on the baselines we established on that matter.

3.4. Dynamic wide coverage lexical resources

Participants: Benoît Sagot, Laurence Danlos, Éric Villemonte de La Clergerie, Marie-Hélène Candito, Lucie Barque, Valérie Hanoka, Marianne Djemaa, Quentin Pradet.

Grammatical formalisms and associated parsing generators are useful only when used together with linguistic resources (lexicons, grammars) so as to build operational parsers, especially when considering modern lexically oriented grammatical formalisms. Hence, linguistic resources are the topic of the following section.

However, wide coverage linguistic resources are scarce and expensive, because they are difficult to build, especially when hand-crafted. This observation motivates us to investigate methods, along to manual development techniques, to automatically or semi-automatically acquire, supplement and correct linguistic resources.

Linguistic expertise remains a very important asset to benefit efficiently from such techniques, including those described below. Moreover, linguistically oriented environments with adequate collaborative interfaces are needed to facilitate the edition, comparison, validation and maintenance of large scale linguistic resources. Just to give some idea of the complexity, a syntactic lexicon, as described below, should provide rich information for several tens of thousands of lemma and several hundreds of thousands of forms.

Successful experiments have been conduced by Alpage members with different languages for the automatic acquisition of morphological knowledge from raw corpora [115]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [142],[6]. At the semantic level, automatic wordnet development tools have been described [104], [137], [84], [81]. All such techniques need of course to be followed by manual validation, so as to ensure high-quality results.

For French, these techniques, and others, have lead some Alpage members to develop one of the main syntactic resources for French, the Le*fff* [111], [117], developed within the Alexina framework. At the semantic level, Alpage members have developed or are developing various syntactico-semantic or semantic resources, including:

- a wordnet for French, the WOLF [112], the first freely available resource of the kind (see 5.7);
- a French FrameNet lexicon (together with an annotated corpus) within the ASFALDA ANR project (see sections 8.1.2.1 and 6.10);
- and a French VerbNet, Verb∋net (see 6.12).

In the last few years, Alpage members have shown how to benefit from other more linguistically-oriented resources, such as the Lexique-Grammaire and DICOVALENCE , in order to improve the coverage and quality of the Le*fff* , the WOLF, the French FrameNet lexicon and the French VerbNet. This work is a good example of how Inria and Paris 7 members of Alpage fruitful collaborate: this collaboration between NLP computer scientists and NLP linguists have resulted in significant advances which would have not been possible otherwise.

Moreover, an increasing effort has been made towards multilingual aspects. In particular, Alexina lexicons developed in 2014 or before exist for German [38], Slovak, Polish, English, Spanish, Persian, Latin (verbs only), Kurmanji Kurdish, Maltese (verbs only, restricted to the so-called first *binyan*) and Khaling, not including freely-available lexicons adapted to the Alexina framework.

3.5. Discourse structures

Participants: Laurence Danlos, James Pustejovsky, Jacques Steinlin, Chloé Braud, Julie Hunter, Raphaël Salmon.

Until now, the linguistic modeling and automatic processing of sentences has been the main focus of the community. However, many applications would benefit from more large-scale approaches which go beyond the level of sentences. This is not only the case for automatic translation: information extraction/retrieval, summarizing, and other applications do need to resolve anaphora, which in turn can benefit from the availability of hierarchical discourse structures induced by discourse relations (in particular through the notion of right frontier of discourse structures). Moreover, discourse structures are required to extract sequential (chronological, logical,...) or hierarchical representations of events. It is also useful for topic extraction, which in turns can help syntactic and semantic disambiguation.

Although supra-sentential problematics received increasing attention in the last years, there is no satisfying solution to these problems. Among them, anaphora resolution and discourse structures have a far-reaching impact and are domains of expertise of Alpage members. But their formal modeling has now reached a maturity which allows to integrate them, in a near future, inside future Alpage tools, including parsing systems inherited from Atoll.

It is well known that a text is not a random sequence of sentences: sentences are linked the ones to the others by "discourse relations", which give to the text a hierarchical structure. Traditionally, it is considered that discourse relations are lexicalized by connectors (adverbial connectors like *ensuite*, conjunctions like *parce que*), or are not lexicalized. This vision is however too simple:

- first, some connectors (in particular conjunctions of subordination) introduce pure modifiers and must not be considered as bearing discourse relations,
- second, other elements than connectors can lexicalize discourse relations, in particular verbs like *précéder / to precede* or *causer / to cause*, which have facts or fact eventualities as arguments [77].

There are three main frameworks used to model discourse structures: RST, SDRT, and, more recently, the TAG-based formalism D-LTAG. Inside Alpage, Laurence Danlos has introduced D-STAG (Discourse Synchronous TAGs, [78],[4]), which subsumes in an elegant way both SDRT and RST, to the extent that SDRT and RST structures can be obtained by two different partial projections of D-STAG structures. As done in D-LTAG, D-STAG extends a lexicalized TAG analysis so as to deal with the level of discourse. D-STAG has been fully formalized, and is hence possible to implement (thanks to Synchronous TAG, or even TAG parsers), provided one develops linguistic descriptions in this formalism.

AVIZ Project-Team

3. Research Program

3.1. Scientific Foundations

The scientific foundations of Visual Analytics lie primarily in the domains of Information Visualization and Data Mining. Indirectly, it inherits from other established domains such as graphic design, Exploratory Data Analysis (EDA), statistics, Artificial Intelligence (AI), Human-Computer Interaction (HCI), and Psychology.

The use of graphic representation to understand abstract data is a goal Visual Analytics shares with Tukey's Exploratory Data Analysis (EDA) [60], graphic designers such as Bertin [45] and Tufte [59], and HCI researchers in the field of Information Visualization [44].

EDA is complementary to classical statistical analysis. Classical statistics starts from a *problem*, gathers *data*, designs a *model* and performs an *analysis* to reach a *conclusion* about whether the data follows the model. While EDA also starts with a problem and data, it is most useful *before* we have a model; rather, we perform visual analysis to discover what kind of model might apply to it. However, statistical validation is not always required with EDA; since often the results of visual analysis are sufficiently clear-cut that statistics are unnecessary.

Visual Analytics relies on a process similar to EDA, but expands its scope to include more sophisticated graphics and areas where considerable automated analysis is required before the visual analysis takes place. This richer data analysis has its roots in the domain of Data Mining, while the advanced graphics and interactive exploration techniques come from the scientific fields of Data Visualization and HCI, as well as the expertise of professions such as cartography and graphic designers who have long worked to create effective methods for graphically conveying information.

The books of the cartographer Bertin and the graphic designer Tufte are full of rules drawn from their experience about how the meaning of data can be best conveyed visually. Their purpose is to find effective visual representation that describe a data set but also (mainly for Bertin) to discover structure in the data by using the right mappings from abstract dimensions in the data to visual ones.

For the last 25 years, the field of Human-Computer Interaction (HCI) has also shown that interacting with visual representations of data in a tight perception-action loop improves the time and level of understanding of data sets. Information Visualization is the branch of HCI that has studied visual representations suitable to understanding and interaction methods suitable to navigating and drilling down on data. The scientific foundations of Information Visualization come from theories about perception, action and interaction.

Several theories of perception are related to information visualization such as the "Gestalt" principles, Gibson's theory of visual perception [51] and Triesman's "preattentive processing" theory [58]. We use them extensively but they only have a limited accuracy for predicting the effectiveness of novel visual representations in interactive settings.

Information Visualization emerged from HCI when researchers realized that interaction greatly enhanced the perception of visual representations.

To be effective, interaction should take place in an interactive loop faster than 100ms. For small data sets, it is not difficult to guarantee that analysis, visualization and interaction steps occur in this time, permitting smooth data analysis and navigation. For larger data sets, more computation should be performed to reduce the data size to a size that may be visualized effectively.

In 2002, we showed that the practical limit of InfoVis was on the order of 1 million items displayed on a screen [48]. Although screen technologies have improved rapidly since then, eventually we will be limited by the physiology of our vision system: about 20 millions receptor cells (rods and cones) on the retina. Another problem will be the limits of human visual attention, as suggested by our 2006 study on change blindness in large and multiple displays [46]. Therefore, visualization alone cannot let us understand very large data sets. Other techniques such as aggregation or sampling must be used to reduce the visual complexity of the data to the scale of human perception.

Abstracting data to reduce its size to what humans can understand is the goal of Data Mining research. It uses data analysis and machine learning techniques. The scientific foundations of these techniques revolve around the idea of finding a good model for the data. Unfortunately, the more sophisticated techniques for finding models are complex, and the algorithms can take a long time to run, making them unsuitable for an interactive environment. Furthermore, some models are too complex for humans to understand; so the results of data mining can be difficult or impossible to understand directly.

Unlike pure Data Mining systems, a Visual Analytics system provides analysis algorithms and processes compatible with human perception and understandable to human cognition. The analysis should provide understandable results quickly, even if they are not ideal. Instead of running to a predefined threshold, algorithms and programs should be designed to allow trading speed for quality and show the tradeoffs interactively. This is not a temporary requirement: it will be with us even when computers are much faster, because good quality algorithms are at least quadratic in time (e.g. hierarchical clustering methods). Visual Analytics systems need different algorithms for different phases of the work that can trade speed for quality in an understandable way.

Designing novel interaction and visualization techniques to explore huge data sets is an important goal and requires solving hard problems, but how can we assess whether or not our techniques and systems provide real improvements? Without this answer, we cannot know if we are heading in the right direction. This is why we have been actively involved in the design of evaluation methods for information visualization [57], [56], [52], [53], [49]. For more complex systems, other methods are required. For these we want to focus on longitudinal evaluation methods while still trying to improve controlled experiments.

3.2. Innovation

We design novel visualization and interaction techniques. Many of these techniques are also evaluated throughout the course of their respective research projects. We cover application domains such as sports analysis, digital humanities, fluid simulations, and biology. A focus of Aviz' work is the improvement of graph visualization and interaction with graphs. We further develop individual techniques for the design of tabular visualizations and different types of data charts. Another focus is the use of animation as a transition aid between different views of the data. We are also interested in applying techniques from illustrative visualization to visual representations and applications in information visualization as well as scientific visualization.

3.3. Evaluation Methods

Evaluation methods are required to assess the effectiveness and usability of visualization and analysis methods. Aviz typically uses traditional HCI evaluation methods, either quantitative (measuring speed and errors) or qualitative (understanding users tasks and activities). Moreover, Aviz is also contributing to the improvement of evaluation methods by reporting on the best practices in the field, by co-organizing workshops (BELIV 2010, 2012, 2014) to exchange on novel evaluation methods, by improving our ways of reporting, interpreting and communicating statistical results, and by applying novel methodologies, for example to assess visualization literacy.

3.4. Software Infrastructures

We want to understand the requirements that software and hardware architectures should provide to support exploratory analysis of large amounts of data. So far, "big data" has been focusing on issues related to storage management and predictive analysis: applying a well-known set of operations on large amounts of data. Visual Analytics is about exploration of data, with sometimes little knowledge of its structure or properties. Therefore, interactive exploration and analysis is needed to build knowledge and apply appropriate analyses; this knowledge and appropriateness is supported by visualizations. However, applying analytical operations on large data implies long-lasting computations, incompatible with interactions, and generates large amounts of results, impossible to visualize directly without aggregation or sampling. Visual Analytics has started to tackle these problems for specific applications but not in a general manner, leading to fragmentation of results and difficulties to reuse techniques from one application to the other. We are interested in abstracting-out the issues and finding general architectural models, patterns, and frameworks to address the Visual Analytics challenge in more generic ways.

3.5. Emerging Technologies

We want to empower humans to make use of data using different types of display media and to enhance how they can understand and visually and interactively explore information. This includes novel display equipment and accompanying input techniques. The Aviz team specifically focuses on the exploration of the use of large displays in visualization contexts as well as emerging physical and tangible visualizations. In terms of interaction modalities our work focuses on using touch and tangible interaction. Aviz participates to the Digiscope project that funds 11 wall-size displays at multiple places in the Paris area (see http://www.digiscope.fr), connected by telepresence equipment and a Fablab for creating devices. Aviz is in charge of creating and managing the Fablab, uses it to create physical visualizations, and is also using the local wall-size display (called WILD) to explore visualization on large screens. The team also investigates the perceptual, motor and cognitive implications of using such technologies for visualization.

3.6. Psychology

More cross-fertilization is needed between psychology and information visualization. The only key difference lies in their ultimate objective: understanding the human mind vs. helping to develop better tools. We focus on understanding and using findings from psychology to inform new tools for information visualization. In many cases, our work also extends previous work in psychology. Our approach to the psychology of information visualization is largely holistic and helps bridge gaps between perception, action and cognition in the context of information visualization. Our focus includes the perception of charts in general, perception in large display environments, collaboration, perception of animations, how action can support perception and cognition, and judgment under uncertainty.

AYIN Team

3. Research Program

3.1. Geometric and shape modeling

One of the grand challenges of computer vision and image processing is the expression and use of prior geometric information via the construction of appropriate models. For very high resolution imagery, this problem becomes critically important, as the increasing resolution of the data results in the appearance of a great deal of complex geometric structure hitherto invisible. Ayin studies various approaches to the construction of models of geometry and shape.

3.1.1. Stochastic geometry

One of the most promising approaches to the inclusion of this type of information is stochastic geometry, which is an important research direction in the Ayin team. Instead of defining probabilities for different types of image, probabilities are defined for configurations of an indeterminate number of interacting, parameterized objects located in the image. Such probability distributions are called 'marked point processes'. New models are being developed both for remote sensing applications, and for skin care problems, such as wrinkle and acne detection.

3.1.2. Contours, phase fields, and MRFs with long-range interactions

An alternative approach to shape modeling starts with generic 'regions' in the image, and adds constraints in order to model specific shapes and objects. Ayin investigates contour, phase field, and binary field representations of regions, incorporating shape information via highly-structured long-range interactions that constrain the set of high-probability regions to those with specific geometric properties. This class of models can represent infinite-dimensional families of shapes and families with unbounded topology, as well as families consisting of an arbitrary number of object instances, at no extra computational cost. Key sub-problems include the development of models of more complex shapes and shape configurations; the development of models in more than two spatial dimensions; and understanding the equivalences between models in different representations and approaches.

3.1.3. Shapes in time

Ayin is concerned with spectral and spatio-temporal structures. To deal with the latter, the above scene modeling approaches are extended into the time dimension, either by modeling time dependence directly, or, in the field-based approaches, by modeling spacetime structures, or, in the stochastic geometry approach, by including the time t in the mark. An example is a spatio-temporal graph-cut-based method that introduces directed infinite links connecting pixels in successive image frames in order to impose constraints on shape change.

3.2. Image modeling

The key issue that arises in modeling the high-resolution image data generated in Ayin's applications, is how to include large-scale spatial, temporal, and spectral dependencies. Ayin investigates approaches to the construction of image models including such dependencies. A central question in teh use of such models is how to deal with the large data volumes arising both from the large size of the images involved, and the existence of large image collections. Fortunately, high dimensionality typically implies data redundancy, and so Ayin investigates methods for reducing the dimensionality of the data and describing the spatial, temporal, and spectral dependencies in ways that allow efficient data processing.

3.2.1. Markov random fields with long-range and higher-order interactions

One way to achieve large-scale dependencies is via explicit long-range interactions. MRFs with long-range interactions are also used in Ayin to model geometric spatial and temporal structure, and the techniques and algorithms developed there will also be applied to image modeling. In modeling image structures, however, other important properties, such as control of the relative phase of Fourier components, and spontaneous symmetry breaking, may also be required. These properties can only be achieved by higher-order interactions. These require specific techniques and algorithms, which are developed in parallel with the models.

3.2.2. Hierarchical models

Another way to achieve long-range dependencies is via shorter range interactions in a hierarchical structure. Ayin works on the development of models defined as a set of hierarchical image partitions represented by a binary forest structure. Key sub-problems include the development of multi-feature models of image regions as an ensemble of spectral, texture, geometrical, and classification features, where we search to optimize the ratio between discrimination capacity of the feature space and dimensionality of this space; and the development of similarity criteria between image regions, which would compute distances between regions in the designed feature space and would be data-driven and scale-independent. One way to proceed in the latter case consists in developing a composite kernel method, which would seek to project multi-feature data into a new space, where regions from different thematic categories become linearly or almost linearly separable. This involves developing kernel functions as a combination of basis kernels, and estimating kernel-based support vector machine parameters.

3.3. Algorithms

Computational techniques are necessary in order to extract the information of interest from the models. In addition, most models contain 'nuisance parameters', including the structure of the models themselves, that must be dealt with in some way. Ayin is interested in adapting and developing methods for solving these problems in cases where existing methods are inadequate.

3.3.1. Nuisance parameters and parameter estimation

In order to render the models operational, it is crucial to find some way to deal with nuisance parameters. In a Bayesian framework, the parameters must be integrated out. Unfortunately, this is usually very difficult. Fortunately, Laplace's method often provides a good approximation, in many cases being equivalent to classical maximum likelihood parameter estimation. Even these problems are not easy to solve, however, when dealing with complex, structured models. This is particularly true when it is necessary to estimate simultaneously both the information of interest and the parameters. Ayin is developing a number of different methods for dealing with nuisance parameters, corresponding to the diversity of modeling approaches.

3.3.2. Information extraction

Extracting the information of interest from any model involves making estimates based on various criteria, for example MAP, MPM, or MMSE. Computing these estimates often requires the solution of hard optimization problems. The complexity of many of the models to be developed within Ayin means that off-the-shelf algorithms and current techniques are often not capable of solving these problems. Ayin develops a diversity of algorithmic approaches adapted to the particular models developed.

DAHU Project-Team

3. Research Program

3.1. Research Program

Dahu aims at developing mechanisms for high-level specifications of systems built around DBMS, that are easy to understand while also facilitating verification of critical properties. This requires developing tools that are suitable for reasoning about systems that manipulate data. Some tools for specifying and reasoning about data have already been studied independently by the database community and by the verification community, with various motivations. However, this work is still in its infancy and needs to be further developed and unified.

Most current proposals for reasoning about DBMS over XML documents are based on tree automata, taking advantage of the tree structure of XML documents. For this reason, the Dahu team is studying a variety of tree automata. This ranges from restrictions of "classical" tree automata in order to understand their expressive power, to extensions of tree automata in order to understand how to incorporate the manipulation of data.

Moreover, Dahu is also interested in logical frameworks that explicitly refer to data. Such logical frameworks can be used as high level declarative languages for specifying integrity constraints, format change during data exchange, web service functionalities and so on. Moreover, the same logical frameworks can be used to express the critical properties we wish to verify.

In order to achieve its goals, Dahu brings together world-class expertise in both databases and verification.

DREAM Project-Team

3. Research Program

3.1. Introduction

The research agenda of the Dream project-team revolves around the following 4 main topics:

- Simulator-based decision support systems
- Incremental learning
- Mining complex patterns
- Answer Set Programming

3.2. Simulator-based decision support systems

A common way to investigate and understand complex phenomena, such as such as those related to ecosystems, consists in designing a computational model and implementing a simulator to test the system behavior under various parameters. These simulators enable a fine grained understanding of the system studied, however they produce huge quantities of data. To be able to exploit these simulators in decision support scenarios, it is thus critical to provide methods to simplify the interactions with the simulator and handle the large quantity of data produced.

- One approach is to store all the simulation data in a datawarehouse and provide scientists and experts with tools to analyze efficiently the simulation data. Providing users with means to dig through large amount of multidimensional data, from more or less abstract viewpoints, and express preferences on the returned results is an important research topic in databases and data mining. To this end, *Skyline queries* constitute a relevant approach as they retrieve the most interesting objects with respect to multi-dimensional criteria with the possibility of making compromises on conflicting dimensions. The challenge is to define and implement skyline queries in a datawarehouse context. In this field, we are investigating efficient interactive tools for answering dynamic [36] and hierarchical [10] skyline queries.
- Another approach is to simplify the simulation model. For some applications, the system is too complex for a traditional numerical simulation to give relevant results in a short amount of time. It is especially the case when data and knowledge are not available to supply numerical models. Qualitative models offers a good alternative to model complex systems in such context. This abstracted representation offers an efficient computation on model exploration and gives relevant results when querying the system behavior. In the Dream project-team we focused on qualitative models of dynamical systems described as Discrete Event Systems (DES). Recent studies have emphasized the great interest of coupling model-checking techniques with qualitative models. We propose to use the timed automata formalism that allow the explicit representation of time [29]. In this context, the research issues we investigate are the following.
  - The size of a global model constructed from an abstracted description of the system and domain knowledge is potentially huge. A challenging problem is to reduce the size of this model using artificial intelligence tools [37].
  - It is necessary to propose a high-level language to explore and predict future changes of the system. Using this language, a stakeholder should express easily any requirements he wants to ask on the system behavior. We investigate the formalization of query patterns relying on recent temporal logics that can be exploited using model-checking techniques [52].

  - Another challenge is the computation of the optimal strategy for a reachability problem (*"what is the best sequence of actions to reach a specific state at a specific time ?"*). In this case we propose to use extended timed automata, such as timed game automata or priced time automata, with controller synthesis methods [30].

- When modelling becomes increasingly complex because of ever-increasing numbers of combined processes, making model-based decision aids are essential. Our approach uses symbolic learning techniques on simulated data to synthetise complex processses and help in decision making. Thus rule induction has attracted a great deal of attention in Machine Learning and Data Mining. However generating rules is not an end in itself because their applicability is not straightforward, especially when their number is high.

  Our goal is to lighten the burden of analyzing a large set of classification rules when the user is confronted to an "unsatisfactory situation" and needs help to decide about the appropriate action to remedy to this situation. The method consists in comparing the situation to a set of classification rules. For this purpose, we have proposed a framework for learning action recommendations dealing with complex notions of feasibility and quality of actions [63].

3.3. Incremental learning

The first learning algorithms were batch learning. They examine all examples and produce a concept description, that is generally not further modified. This is not adapted to dynamic settings where data are delivered continuously. For such settings, incremental algorithms have been proposed. These algorithms examine the training example one at a time (or set by set), maintaining a "best-so-far" description which may be modified each time a new example (or set of examples) arrives. In order to strengthen the learning process, some specific old examples are often kept: this is called partial memory systems. A more specific classification of incremental learning can be found in [58].

Current issues in incremental learning are

- for partial instance memory: how to select examples, [56]
- the problem of hidden context: the target concept may depend of unknown variables, which are not given as explicit attributes [66]
- the problem of concept drift: the target changes with time [65], [39]
- the problem of masked example: the data distribution may change and some examples may not be anymore visible.

As a human expert has to give his opinion on the learned description model, we focus our research on incremental learning of rules ([39]).

3.4. Mining complex patterns

Pattern mining, a subdomain of data mining, is an unsupervised learning method which aims at discovering interesting knowledge from data. Association rule extraction is one of the most popular approach and has received a lot of interest in the last 20 years. For instance, many enhancements have been proposed to the well-known Apriori algorithm [27]. It is based on a level-wise generation of candidate patterns and on efficient candidate pruning having a sufficient relevance, usually related to the frequency of the candidate pattern in the data-set (i.e., the support): the most frequent patterns should be the most interesting. Later, Agrawal and Srikant proposed a framework for "mining sequential patterns" [28], which extends Apriori by coping with the order of elements in patterns. Such approach initiated research on *temporal pattern mining*, which is of particular interest for the DREAM team. The simplest temporal patterns are sequential patterns that constraints the order of the events in one of its occurrence. More advanced approaches also exploit quantitative information in order to provide significant patterns about both ordering and duration of events as well as inter-event delay. A challenge is that the classical anti-monotony property, used to prune the search space, is difficult to define in this case.

Many work in pattern mining have attempted to improve the runtime efficiency of algorithms, on the one hand, by proposing more efficient representation and execution schemes such as pattern-growth methods [48], or, on the other hand, by focusing on condensed representations such as closed patterns [60], [64]. Other research directions have been investigated to enhance the syntax of patterns e.g. temporal and periodic patterns, mutidimensional and hierarchical patterns, constrained patterns, contextual patterns, etc. Despite these improvements, the size of the results may still be too high. Thus, post-mining or visualization methods have been introduced to let the user focus on results that correspond to his own preferences.

Another challenge of pattern mining is that for each pattern mining task (such as mining itemsets, sequences or graphs) there are many specialized algorithms, each exploiting some ad-hoc optimizations. It is very hard for a practitioner to find an algorithm suited for his problem, and such an algorithm may not exist. There is a need to propose novel *generic* pattern mining algorithms, that exploit the main algorithmic advances proposed in the last 20 years, and that only require a description of their pattern mining problem from practioners. Recently, we have proposed ParaMiner [59], a generic pattern mining algorithm using state of the art optimizations and exploiting the parallelism of multicore processors. The practitioner only has to enter a pattern interest criteria and check that it verifies a *strong accessibility* property coming from set theory. As of now, ParaMiner is the fastest generic pattern mining algorithm, being competitive with specialized algorithm on several pattern mining tasks.

Other approaches propose a completely declarative way to specify the pattern mining problem. In this case, the most used framework is Constraint Programming [44]. We are investigating another approach based on *Answer Set Programming*.

3.5. Answer Set Programing (ASP)

The DREAM team is investigating declarative approaches to solve complex problems such as causal reasoning, landscape simultation and pattern mining. One such approach is ASP.

ASP (Answer set programming) [43], [31] is an approach to declarative problem solving, combining a rich yet simple modelling language with high-performance solving capacities, tailored to Knowledge Representation and Reasoning. "Declarative problem solving" means that the program is close to the way a problem is enunciated, and not to the way the problem is solved. This facilitates writing and revising programs. ASP is an outgrowth of research on the use of non monotonic reasoning in knowledge representation. ASP programs[23] consist in rules that look like Prolog rules, but the computational mechanism is different [54].

ASP allows to solve search problems in NP (and theoretically in NP^{NP}) in a uniform way (being more compact than boolean approaches like SAT solvers). ASP is good when dealing with knowledge representation, particularly when logical rules or graphs are involved. The versatility of ASP is reflected by the ASP solver clasp, winning first places at ASP, SAT and other competitions.

ASP solvers deal with propositional rules, however in practice predicates are allowed. A *grounder* replaces each free variable of the program provided by users with any eligible constant symbol. The output of the grounder is thus a propositional program, which is piped into a *solver* which then computes *answer sets*. These answer sets are the models for the ASP theory, and they constitute the result of an ASP program. The user may ask for all the models, or only one, or any number n of models. The most powerful version (clingo, which combines the grounder gringo and the solver clasp) is from Torsten Schaub's team (see http://potassco.sourceforge.net for the last version of clingo, including a *guide*). These versions can be easily interfaced with python programs, which extends further the practical applicability of ASP [42].

The main interests of using ASP are: 1) the ease to write and to update programs, and 2) the efficiency of the ASP solvers (improved in the recent versions).

Our main challenge is to propose ASP modeling that scales up to solving real problems. We are especially working on the modeling of sequential pattern mining with ASP in order to mine real datasets in a flexible and efficient way.

Our second challenge is to model a wide range of expert knowledge to include reasoning into the solving processes, in order to output more meaningful results.

E-MOTION Project-Team (section vide)

EXMO Project-Team

3. Research Program

3.1. Knowledge representation semantics

We usually work with semantically defined knowledge representation languages (like description logics, conceptual graphs and object-based languages) [16]. Their semantics is usually defined within model theory initially developed for logics. The languages dedicated to the semantic web (RDF and OWL) follow that approach. RDF is a knowledge representation language dedicated to the description of resources; OWL is designed for expressing ontologies: it describes concepts and relations that can be used within RDF.

We consider a language L as a set of syntactically defined expressions (often inductively defined by applying constructors over other expressions). A representation ($o \subseteq L$) is a set of such expressions. It is also called an ontology. An interpretation function (I) is inductively defined over the structure of the language to a structure called interpretation domain (D). This expresses the construction of the "meaning" of an expression in function of its components. A formula is satisfied by an interpretation if it fulfills a condition (in general being interpreted over a particular subset of the domain). A model of a set of expressions is an interpretation satisfying all these expressions. An expression (δ) is then a consequence of a set of expressions (o) if it is satisfied by all of their models (noted $o \models \delta$).

A computer must determine if a particular expression (taken as a query, for instance) is the consequence of a set of axioms (a knowledge base). For that purpose, it uses programs, called provers, that can be based on the processing of a set of inference rules, on the construction of models or on procedural programming. These programs are able to deduce theorems (noted $o \vdash \delta$). They are said to be sound if they only find theorems which are indeed consequences and to be complete if they find all the consequences as theorems. However, depending on the language and its semantics, the decidability, i.e., the ability to create sound and complete provers, is not warranted. Even for decidable languages, the algorithmic complexity of provers may prohibit their exploitation.

To solve this problem a trade-off between the expressivity of the language and the complexity of its provers has to be found. These considerations have led to the definition of languages with limited complexity – like conceptual graphs and object-based representations – or of modular families of languages with associated modular prover algorithms – like description logics.

EXMO mainly considers languages with well-defined semantics (such as RDF and OWL that we contributed to define), and defines the semantics of some languages such as the SPARQL query language and alignment languages, in order to establish the properties of computer manipulations of the representations.

3.2. Ontology matching and alignments

When different representations are used, it is necessary to identify their correspondences. This task is called ontology matching and its result is an alignment [3]. It can be described as follows: given two ontologies, each describing a set of discrete entities (which can be classes, properties, rules, predicates, etc.), find the relationships, e.g., equivalence or subsumption, if any, holding between these entities.

An alignment between two ontologies o and o' is a set of correspondences $\langle e, e', r \rangle$ such that:

- e and e' are the entities between which a relation is asserted by the correspondence, e.g., formulas, terms, classes, individuals;
- r is the relation asserted to hold between e and e'. This relation can be any relation applying to these entities, e.g., equivalence, subsumption.

In addition, a correspondence may support various types of metadata, in particular measures of the confidence in a correspondence.

Given the semantics of the two ontologies provided by their consequence relation, we define an interpretation of two aligned ontologies as a pair of interpretations $\langle m, m' \rangle$, one for each ontology. Such a pair of interpretations is a model of the aligned ontologies o and o' if and only if each respective interpretation is a model of the ontology and they satisfy all correspondences of the alignment.

This definition is extended to networks of ontologies: a collection of ontologies and associated alignments. A model of such an ontology network is a tuple of local models such that each alignment is valid for the models involved in the tuple. In such a system, alignments play the role of model filters which select the local models that are compatible with all alignments. So, given an ontology network, it is possible to interpret it.

However, given a set of ontologies, it is necessary to find the alignments between them and the semantics does not tell which ones they are. Ontology matching aims at finding these alignments. A variety of methods is used for this task. They perform pairwise comparisons of entities from each of the ontologies and select the most similar pairs. Most matching algorithms provide correspondences between named entities, more rarely between compound terms. The relationships are generally equivalence between these entities. Some systems are able to provide subsumption relations as well as other relations in the support language (like incompatibility or instantiation). Confidence measures are usually given a value between 0 and 1 and are used for expressing preferences between two correspondences.

3.3. Data interlinking

Links are important for the publication of RDF data on the web. We call data interlinking the process of generating links identifying same resource described in two data sets. Data interlinking parallels ontology matching: from two datasets (d and d') it generates a set of links (also called a linkset, L).

We have extended the notion of database keys in a way which is more adapted to the context of description logics and the openness of the semantic web [11] [0]. Like alignments, link keys [3] are assertions across ontologies and are not part of a single ontology. We have introduced the notion of a link key which is a combination of such keys with alignments. More precisely, a link key is an expression $\langle K^{eq}, K^{in}, C \rangle$ such that:

- K^{eq} is a set of pairs of property expressions;
- K^{in} is a set of pairs of property expressions;
- C is a correspondence between classes.

Such a link key holds if and only if for any pair of resources belonging to the classes in correspondence such that the values of their property in K^{eq} are pairwise equal and the values of those in K^{in} pairwise intersect, the resources are the same.

As can be seen, link key validity is only relying on pairs of objects in two different data sets. We further qualify link keys as weak, plain and strong depending on them satisfying further constraints: a weak link key is only valid on pairs of individuals of different data sets, a plain link key has to apply in addition to pairs of individuals of the same data set as soon as one of them is identified with another individual of the other data set, a strong link key is a link key which is also a key for each data set, it can be though of as a link key which is made of two keys.

Link keys can then be used for finding equal individuals across the two data sets and generating the corresponding owl:sameAs links.

[0] Time did not permit to input properly all publications in HAL v3. We understand well that these are thus not Inria publications. However, we put them as footnotes in case they may interest the reader. They are all directly available from our team web site.

FLOWERS Project-Team

3. Research Program

3.1. Research Program

Research in artificial intelligence, machine learning and pattern recognition has produced a tremendous amount of results and concepts in the last decades. A blooming number of learning paradigms - supervised, unsupervised, reinforcement, active, associative, symbolic, connectionist, situated, hybrid, distributed learning... - nourished the elaboration of highly sophisticated algorithms for tasks such as visual object recognition, speech recognition, robot walking, grasping or navigation, the prediction of stock prices, the evaluation of risk for insurances, adaptive data routing on the internet, etc... Yet, we are still very far from being able to build machines capable of adapting to the physical and social environment with the flexibility, robustness, and versatility of a one-year-old human child.

Indeed, one striking characteristic of human children is the nearly open-ended diversity of the skills they learn. They not only can improve existing skills, but also continuously learn new ones. If evolution certainly provided them with specific pre-wiring for certain activities such as feeding or visual object tracking, evidence shows that there are also numerous skills that they learn smoothly but could not be "anticipated" by biological evolution, for example learning to drive a tricycle, using an electronic piano toy or using a video game joystick. On the contrary, existing learning machines, and robots in particular, are typically only able to learn a single pre-specified task or a single kind of skill. Once this task is learnt, for example walking with two legs, learning is over. If one wants the robot to learn a second task, for example grasping objects in its visual field, then an engineer needs to re-program manually its learning structures: traditional approaches to task-specific machine/robot learning typically include engineer choices of the relevant sensorimotor channels, specific design of the reward function, choices about when learning begins and ends, and what learning algorithms and associated parameters shall be optimized.

As can be seen, this requires a lot of important choices from the engineer, and one could hardly use the term "autonomous" learning. On the contrary, human children do not learn following anything looking like that process, at least during their very first years. Babies develop and explore the world by themselves, focusing their interest on various activities driven both by internal motives and social guidance from adults who only have a folk understanding of their brains. Adults provide learning opportunities and scaffolding, but eventually young babies always decide for themselves what activity to practice or not. Specific tasks are rarely imposed to them. Yet, they steadily discover and learn how to use their body as well as its relationships with the physical and social environment. Also, the spectrum of skills that they learn continuously expands in an organized manner: they undergo a developmental trajectory in which simple skills are learnt first, and skills of progressively increasing complexity are subsequently learnt.

A link can be made to educational systems where research in several domains have tried to study how to provide a good learning experience to learners. This includes the experiences that allow better learning, and in which sequence they must be experienced. This problem is complementary to that of the learner that tries to learn efficiently, and the teacher here has to use as efficienlty the limited time and motivational resources of the learner. Several results from psychology [76] and neuroscience [10] have argued that the human brain feels intrinsic pleasure in practicing activities of optimal difficulty or challenge. A teacher must exploit such activities to create positive psychological states of flow [82].

A grand challenge is thus to be able to build robotic machines that possess this capability to discover, adapt and develop continuously new know-how and new knowledge in unknown and changing environments, like human children. In 1950, Turing wrote that the child's brain would show us the way to intelligence: "Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's" [116]. Maybe, in opposition to work in the field of Artificial Intelligence who has focused on mechanisms trying to match the capabilities of "intelligent" human adults such as chess playing or natural

language dialogue [89], it is time to take the advice of Turing seriously. This is what a new field, called developmental (or epigenetic) robotics, is trying to achieve [96] [118]. The approach of developmental robotics consists in importing and implementing concepts and mechanisms from developmental psychology [101], cognitive linguistics [81], and developmental cognitive neuroscience [93] where there has been a considerable amount of research and theories to understand and explain how children learn and develop. A number of general principles are underlying this research agenda: embodiment [78] [107], grounding [87], situatedness [72], self-organization [114] [108], enaction [117], and incremental learning [79].

Among the many issues and challenges of developmental robotics, two of them are of paramount importance: exploration mechanisms and mechanisms for abstracting and making sense of initially unknown sensorimotor channels. Indeed, the typical space of sensorimotor skills that can be encountered and learnt by a developmental robot, as those encountered by human infants, is immensely vast and inhomogeneous. With a sufficiently rich environment and multimodal set of sensors and effectors, the space of possible sensorimotor activities is simply too large to be explored exhaustively in any robot's life time: it is impossible to learn all possible skills and represent all conceivable sensory percepts. Moreover, some skills are very basic to learn, some other very complicated, and many of them require the mastery of others in order to be learnt. For example, learning to manipulate a piano toy requires first to know how to move one's hand to reach the piano and how to touch specific parts of the toy with the fingers. And knowing how to move the hand might require to know how to track it visually.

Exploring such a space of skills randomly is bound to fail or result at best on very inefficient learning [15]. Thus, exploration needs to be organized and guided. The approach of epigenetic robotics is to take inspiration from the mechanisms that allow human infants to be progressively guided, i.e. to develop. There are two broad classes of guiding mechanisms which control exploration:

1. **internal guiding mechanisms,** and in particular intrinsic motivation, responsible of spontaneous exploration and curiosity in humans, which is one of the central mechanisms investigated in FLOWERS, and technically amounts to achieve online active self-regulation of the growth of complexity in learning situations;
2. **social learning and guidance,** a learning mechanisms that exploits the knowledge of other agents in the environment and/or that is guided by those same agents. These mechanisms exist in many different forms like emotional reinforcement, stimulus enhancement, social motivation, guidance, feedback or imitation, some of which being also investigated in FLOWERS;

3.1.1. Internal guiding mechanisms

In infant development, one observes a progressive increase of the complexity of activities with an associated progressive increase of capabilities [101], children do not learn everything at one time: for example, they first learn to roll over, then to crawl and sit, and only when these skills are operational, they begin to learn how to stand. The perceptual system also gradually develops, increasing children perceptual capabilities other time while they engage in activities like throwing or manipulating objects. This make it possible to learn to identify objects in more and more complex situations and to learn more and more of their physical characteristics.

Development is therefore progressive and incremental, and this might be a crucial feature explaining the efficiency with which children explore and learn so fast. Taking inspiration from these observations, some roboticists and researchers in machine learning have argued that learning a given task could be made much easier for a robot if it followed a developmental sequence and "started simple" [74] [85]. However, in these experiments, the developmental sequence was crafted by hand: roboticists manually build simpler versions of a complex task and put the robot successively in versions of the task of increasing complexity. And when they wanted the robot to learn a new task, they had to design a novel reward function.

Thus, there is a need for mechanisms that allow the autonomous control and generation of the developmental trajectory. Psychologists have proposed that intrinsic motivations play a crucial role. Intrinsic motivations are mechanisms that push humans to explore activities or situations that have intermediate/optimal levels of novelty, cognitive dissonance, or challenge [76] [82] [84]. The role and structure of intrinsic motivation in humans have been made more precise thanks to recent discoveries in neuroscience showing the implication

of dopaminergic circuits and in exploration behaviors and curiosity [83] [90] [113]. Based on this, a number of researchers have began in the past few years to build computational implementation of intrinsic motivation [15] [105] [111] [75] [91] [99] [112]. While initial models were developed for simple simulated worlds, a current challenge is to manage to build intrinsic motivation systems that can efficiently drive exploratory behaviour in high-dimensional unprepared real world robotic sensorimotor spaces [105][15] [106] [110]. Specific and complex problems are posed by real sensorimotor spaces, in particular due to the fact that they are both high-dimensional as well as (usually) deeply inhomogeneous. As an example for the latter issue, some regions of real sensorimotor spaces are often unlearnable due to inherent stochasticity or difficulty, in which case heuristics based on the incentive to explore zones of maximal unpredictability or uncertainty, which are often used in the field of active learning [80] [88] typically lead to catastrophic results. The issue of high dimensionality does not only concern motor spaces, but also sensory spaces, leading to the problem of correctly identifying, among typically thousands of quantities, those latent variables that have links to behavioral choices. In FLOWERS, we aim at developing intrinsically motivated exploration mechanisms that scale in those spaces, by studying suitable abstraction processes in conjunction with exploration strategies.

3.1.2. Socially Guided and Interactive Learning

Social guidance is as important as intrinsic motivation in the cognitive development of human babies [101]. There is a vast literature on learning by demonstration in robots where the actions of humans in the environment are recognized and transferred to robots [73]. Most such approaches are completely passive: the human executes actions and the robot learns from the acquired data. Recently, the notion of interactive learning has been introduced in [115], [77], motivated by the various mechanisms that allow humans to socially guide a robot [109]. In an interactive context the steps of self-exploration and social guidances are not separated and a robot learns by self exploration and by receiving extra feedback from the social context [115], [94] [100].

Social guidance is also particularly important for learning to segment and categorize the perceptual space. Indeed, parents interact a lot with infants, for example teaching them to recognize and name objects or characteristics of these objects. Their role is particularly important in directing the infant attention towards objects of interest that will make it possible to simplify at first the perceptual space by pointing out a segment of the environment that can be isolated, named and acted upon. These interactions will then be complemented by the children own experiments on the objects chosen according to intrinsic motivation in order to improve the knowledge of the object, its physical properties and the actions that could be performed with it.

In FLOWERS, we are aiming at including intrinsic motivation system in the self-exploration part thus combining efficient self-learning with social guidance [103], [104]. We also work on developing perceptual capabilities by gradually segmenting the perceptual space and identifying objects and their characteristics through interaction with the user [97] and robots experiments [92]. Another challenge is to allow for more flexible interaction protocols with the user in terms of what type of feedback is provided and how it is provided [95].

Exploration mechanisms are combined with research in the following directions:

3.1.3. Cumulative learning, reinforcement learning and optimization of autonomous skill learning

FLOWERS develops machine learning algorithms that can allow embodied machines to acquire cumulatively sensorimotor skills. In particular, we develop optimization and reinforcement learning systems which allow robots to discover and learn dictionaries of motor primitives, and then combine them to form higher-level sensorimotor skills.

3.1.4. Autonomous perceptual and representation learning

In order to harness the complexity of perceptual and motor spaces, as well as to pave the way to higher-level cognitive skills, developmental learning requires abstraction mechanisms that can infer structural information out of sets of sensorimotor channels whose semantics is unknown, discovering for example the topology of the body or the sensorimotor contingencies (proprioceptive, visual and acoustic). This process is meant to

be open- ended, progressing in continuous operation from initially simple representations towards abstract concepts and categories similar to those used by humans. Our work focuses on the study of various techniques for:

- autonomous multimodal dimensionality reduction and concept discovery;
- incremental discovery and learning of objects using vision and active exploration, as well as of auditory speech invariants;
- learning of dictionaries of motion primitives with combinatorial structures, in combination with linguistic description;
- active learning of visual descriptors useful for action (e.g. grasping);

3.1.5. Embodiment and maturational constraints

FLOWERS studies how adequate morphologies and materials (i.e. morphological computation), associated to relevant dynamical motor primitives, can importantly simplify the acquisition of apparently very complex skills such as full-body dynamic walking in biped. FLOWERS also studies maturational constraints, which are mechanisms that allow for the progressive and controlled release of new degrees of freedoms in the sensorimotor space of robots.

3.1.6. Discovering and abstracting the structure of sets of uninterpreted sensors and motors

FLOWERS studies mechanisms that allow a robot to infer structural information out of sets of sensorimotor channels whose semantics is unknown, for example the topology of the body and the sensorimotor contingencies (propriocetive, visual and acoustic). This process is meant to be open-ended, progressing in continuous operation from initially simple representations to abstract concepts and categories similar to those used by humans.

GRAPHIK Project-Team

3. Research Program

3.1. Logic-based Knowledge Representation and Reasoning

We follow the mainstream *logic-based* approach to the KRR domain. First-order logic (FOL) is the reference logic in KRR and most formalisms in this area can be translated into fragments (i.e., particular subsets) of FOL. A large part of research in this domain can be seen as studying the *trade-off* between the expressivity of languages and the complexity of (sound and complete) reasoning in these languages. The fundamental problem in KRR languages is entailment checking: is a given piece of knowledge entailed by other pieces of knowledge, for instance from a knowledge base (KB)? Another important problem is *consistency* checking: is a set of knowledge pieces (for instance the knowledge base itself) consistent, i.e., is it sure that nothing absurd can be entailed from it? The *ontological query answering* problem is a topical problem (see Section 3.3). It asks for the set of answers to a query in the KB. In the case of Boolean queries (i.e., queries with a yes/no answer), it can be recast as entailment checking.

3.2. Graph-based Knowledge Representation and Reasoning

Besides logical foundations, we are interested in KRR formalisms that comply, or aim at complying with the following requirements: to have good *computational* properties and to allow users of knowledge-based systems to have a maximal *understanding and control* over each step of the knowledge base building process and use.

These two requirements are the core motivations for our specific approach to KRR, which is based on labelled *graphs*. Indeed, we view labelled graphs as an *abstract representation* of knowledge that can be expressed in many KRR languages (different kinds of conceptual graphs —historically our main focus—, the Semantic Web language RDF (Resource Description Framework), its extension RDFS (RDF Schema), expressive rules equivalent to the so-called tuple-generating-dependencies in databases, some description logics dedicated to query answering, etc.). For these languages, reasoning can be based on the structure of objects, thus based on graph-theoretic notions, while staying logically founded.

More precisely, our basic objects are labelled graphs (or hypergraphs) representing entities and relationships between these entities. These graphs have a natural translation in first-order logic. Our basic reasoning tool is graph homomorphism. The fundamental property is that graph homomorphism is sound and complete with respect to logical entailment *i.e.*, given two (labelled) graphs G and H, there is a homomorphism from G to H *if and only if* the formula assigned to G is entailed by the formula assigned to H. In other words, logical reasoning on these graphs can be performed by graph mechanisms. These knowledge constructs and the associated reasoning mechanisms can be extended (to represent rules for instance) while keeping this fundamental correspondence between graphs and logics.

3.3. Ontological Query Answering

Querying knowledge bases has become a central problem in knowledge representation and in databases. A knowledge base (KB) is classically composed of a terminological part (metadata, ontology) and an assertional part (facts, data). Queries are supposed to be at least as expressive as the basic queries in databases, i.e., conjunctive queries, which can be seen as existentially closed conjunctions of atoms or as labelled graphs. The challenge is to define good trade-offs between the expressivity of the ontological language and the complexity of querying data in presence of ontological knowledge. Classical ontogical languages, typically description logics, were not designed for efficient querying. On the other hand, database languages are able to process complex queries on huge databases, but without taking the ontology into account. There is thus a need for new languages and mechanisms, able to cope with the ever growing size of knowledge bases in the Semantic Web or in scientific domains.

This problem is related to two other problems identified as fundamental in KRR:

- *Query-answering with incomplete information.* Incomplete information means that it might be unknown whether a given assertion is true or false. Databases classically make the so-called closed-world assumption: every fact that cannot be retrieved or inferred from the base is assumed to be false. Knowledge bases classically make the open-world assumption: if something cannot be inferred from the base, and neither can its negation, then its truth status is unknown. The need of coping with incomplete information is a distinctive feature of querying knowledge bases with respect to querying classical databases (however, as explained above, this distinction tends to disappear). The presence of incomplete information makes the query answering task much more difficult.
- *Reasoning with rules.* Researching types of rules and adequate manners to process them is a mainstream topic in the Semantic Web, and, more generally a crucial issue for knowledge-based systems. For several years, we have been studying some rules, both in their logical and their graph form, which are syntactically very simple but also very expressive. These rules, known as existential rules or Datalog$^+$, can be seen as an abstraction of ontological knowledge expressed in the main languages used in the context of KB querying. See Section 6.2 for details on the results obtained.

A problem generalizing the above described problems, and particularly relevant in the context of multiple data/metadata sources, is *querying hybrid knowledge bases*. In a hybrid knowledge base, each component may have its own formalism and its own reasoning mechanisms. There may be a common ontology shared by all components, or each component may have its own ontology, with mappings being defined among the ontologies. The question is what kind of interactions between these components and/or what limitations on the languages preserve the decidability of basic problems and if so, a "reasonable"complexity. Note that there are strong connections with the issue of data integration in databases.

3.4. Imperfect Information and Priorities

While classical FOL is the kernel of many KRR languages, to solve real-world problems we often need to consider features that cannot be expressed purely (or not naturally) in classical logic. The logic- and graph-based formalisms used for previous points have thus to be extended with such features.The following requirements have been identified from scenarios in decision making in the agronomy domain (see Section 4.2):

1. to cope with vague and uncertain information and preferences in queries;
2. to cope with multi-granularity knowledge;
3. to take into account different and potentially conflicting viewpoints ;
4. to integrate decision notions (priorities, gravity, risk, benefit);
5. to integrate argumentation-based reasoning.

Although the solutions we develop need to be validated on the applications that motivated them, we also want them to be sufficiently generic to be applied in other contexts. One angle of attack (but not the only possible one) consists in increasing the expressivity of our core languages, while trying to preserve their essential combinatorial properties, so that algorithmic optimizations can be transferred to these extensions. To achieve that goal, our main research directions are: non-monotonic reasoning (see ANR project ASPIQ in Section 8.1), as well as argumentation and preferences (see Section 6.3).

HEPHAISTOS Team

3. Research Program

3.1. Interval analysis

We are interested in real-valued system solving ($f(X) = 0$, $f(X) \leq 0$), in optimization problems, and in the proof of the existence of properties (for example, it exists X such that $f(X) = 0$ or it exist two values X_1, X_2 such that $f(X_1) > 0$ and $f(X_2) < 0$). There are few restrictions on the function f as we are able to manage explicit functions using classical mathematical operators (e.g. $\sin(x + y) + \log(\cos(e^x) + y^2)$ as well as implicit functions (e.g. determining if there are parameter values of a parametrized matrix such that the determinant of the matrix is negative, without calculating the analytical form of the determinant).

Solutions are searched within a finite domain (called a *box*) which may be either continuous or mixed (i.e. for which some variables must belong to a continuous range while other variables may only have values within a discrete set). An important point is that we aim at finding all the solutions within the domain whenever the computer arithmetic will allow it: in other words we are looking for *certified* solutions. For example, for 0-dimensional system solving, we will provide a box that contains one, and only one, solution together with a numerical approximation of this solution. This solution may further be refined at will using multi-precision.

The core of our methods is the use of *interval analysis* that allows one to manipulate mathematical expressions whose unknowns have interval values. A basic component of interval analysis is the *interval evaluation* of an expression. Given an analytical expression F in the unknowns $\{x_1, x_2, ..., x_n\}$ and ranges $\{X_1, X_2, ..., X_n\}$ for these unknowns we are able to compute a range $[A, B]$, called the interval evaluation, such that

$$\forall \{x_1, x_2, ..., x_n\} \in \{X_1, X_2, ..., X_n\}, A \leq F(x_1, x_2, ..., x_n) \leq B \tag{90}$$

In other words the interval evaluation provides a lower bound of the minimum of F and an upper bound of its maximum over the box.

For example if $F = x\ sin(x + x^2)$ and $x \in [0.5, 1.6]$, then $F([0.5, 1.6]) = [-1.362037441, 1.6]$, meaning that for any x in [0.5,0.6] we guarantee that $-1.362037441 \leq f(x) \leq 1.6$.

The interval evaluation of an expression has interesting properties:

- it can be implemented in such a way that the results are guaranteed with respect to round-off errors i.e. property 1 is still valid in spite of numerical errors induced by the use of floating point numbers
- if $A > 0$ or $B < 0$, then no values of the unknowns in their respective ranges can cancel F
- if $A > 0$ ($B < 0$), then F is positive (negative) for any value of the unknowns in their respective ranges

A major drawback of the interval evaluation is that $A(B)$ may be overestimated i.e. values of $x_1, x_2, ..., x_n$ such that $F(x_1, x_2, ..., x_n) = A(B)$ may not exist. This overestimation occurs because in our calculation each occurrence of a variable is considered as an independent variable. Hence if a variable has multiple occurrences, then an overestimation may occur. Such phenomena can be observed in the previous example where $B = 1.6$ while the real maximum of F is approximately 0.9144. The value of B is obtained because we are using in our calculation the formula $F = xsin(y + z^2)$ with y, z having the same interval value than x.

Fortunately there are methods that allow one to reduce the overestimation and the overestimation amount decreases with the width of the ranges. The latter remark leads to the use of a branch-and-bound strategy in which for a given box a variable range will be bisected, thereby creating two new boxes that are stored in a list and processed later on. The algorithm is complete if all boxes in the list have been processed, or if during the process a box generates an answer to the problem at hand (e.g. if we want to prove that $F(X) < 0$, then the algorithm stops as soon as $F(\mathcal{B}) \geq 0$ for a certain box $\mathcal{B}$).

A generic interval analysis algorithm involves the following steps on the current box [1], [8], [5]:

1. *exclusion operators*: these operators determine that there is no solution to the problem within a given box. An important issue here is the extensive and smart use of the monotonicity of the functions
2. *filters*: these operators may reduce the size of the box i.e. decrease the width of the allowed ranges for the variables
3. *existence operators*: they allow one to determine the existence of a unique solution within a given box and are usually associated with a numerical scheme that allows for the computation of this solution in a safe way
4. *bisection*: choose one of the variable and bisect its range for creating two new boxes
5. *storage*: store the new boxes in the list

The scope of the HEPHAISTOS project is to address all these steps in order to find the most efficient procedures. Our efforts focus on mathematical developments (adapting classical theorems to interval analysis, proving interval analysis theorems), the use of symbolic computation and formal proofs (a symbolic pre-processing allows one to automatically adapt the solver to the structure of the problem), software implementation and experimental tests (for validation purposes).

3.2. Robotics

HEPHAISTOS, as a follow-up of COPRIN, has a long-standing tradition of robotics studies, especially for closed-loop robots [4], especially cable-driven parallel robots. We address theoretical issues with the purpose of obtaining analytical and theoretical solutions, but in many cases only numerical solutions can be obtained due to the complexity of the problem. This approach has motivated the use of interval analysis for two reasons:

1. the versatility of interval analysis allows us to address issues (e.g. singularity analysis) that cannot be tackled by any other method due to the size of the problem
2. uncertainties (which are inherent to a robotic device) have to be taken into account so that the *real* robot is guaranteed to have the same properties as the *theoretical* one, even in the worst case. This is a crucial issue for many applications in robotics (e.g. medical or assistance robot)

Our field of study in robotics focuses on *kinematic* issues such as workspace and singularity analysis, positioning accuracy, trajectory planning, reliability, calibration, modularity management and, prominently, *appropriate design*, i.e. determining the dimensioning of a robot mechanical architecture that guarantees that the real robot satisfies a given set of requirements. The methods that we develop can be used for other robotic problems, see for example the management of uncertainties in aircraft design [6].

Our theoretical work must be validated through experiments that are essential for the sake of credibility. A contrario, experiments will feed theoretical work. Hence HEPHAISTOS works with partners on the development of real robots but also develops its own prototypes. In the last years we have developed a large number of prototypes and we have extended our development to devices that are not strictly robots but are part of an overall environment for assistance. We benefit here from the development of new miniature, low energy computers with an interface for analog and logical sensors such as the Arduino or the Phidgets.

HYBRID Project-Team

3. Research Program

3.1. Research Program

The scientific objective of Hybrid team is to improve 3D interaction of one or multiple users with virtual environments, by making full use of physical engagement of the body, and by incorporating the mental states by means of brain-computer interfaces. We intend to improve each component of this framework individually, but we also want to improve the subsequent combinations of these components.

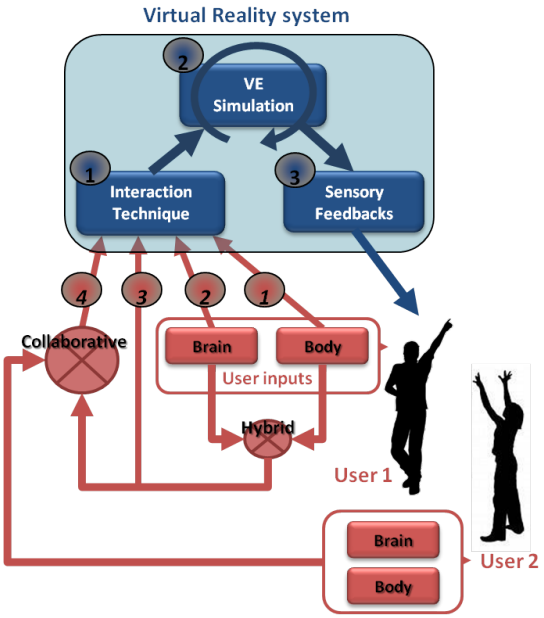


Figure 1. 3D hybrid interaction loop between one or multiple users and a virtual reality system. Top (in blue) three steps of 3D interaction with a virtual environment: (1) interaction technique, (2) simulation of the virtual environment, (3) sensory feedbacks. Bottom (in red) different cases of interaction: (1) body-based, (2) mind-based, (3) hybrid, and (4) collaborative 3D interaction.

The "hybrid" 3D interaction loop between one or multiple users and a virtual environment is depicted on Figure 1 . Different kinds of 3D interaction situations are distinguished (red arrows, bottom): 1) body-based interaction, 2) mind-based interaction, 3) hybrid and/or 4) collaborative interaction (with at least two users). In each case, three scientific challenges arise which correspond to the three successive steps of the 3D interaction loop (blue squares, top): 1) the 3D interaction technique, 2) the modeling and simulation of the 3D scenario, and 3) the design of appropriate sensory feedback.

The 3D interaction loop involves various possible inputs from the user(s) and different kinds of output (or sensory feedback) from the simulated environment. Each user can involve his/her body and mind by means of corporal and/or brain-computer interfaces. A hybrid 3D interaction technique (1) mixes mental and motor inputs and translates them into a command for the virtual environment. The real-time simulation (2) of the

virtual environment is taking into account these commands to change and update the state of the virtual world and virtual objects. The state changes are sent back to the user and perceived by means of different sensory feedbacks (e.g., visual, haptic and/or auditory) (3). The sensory feedbacks are closing the 3D interaction loop. Other users can also interact with the virtual environment using the same procedure, and can eventually "collaborate" by means of "collaborative interactive techniques" (4).

This description is stressing three major challenges which correspond to three mandatory steps when designing 3D interaction with virtual environments:

- **3D interaction techniques:** This first step consists in translating the actions or intentions of the user (inputs) into an explicit command for the virtual environment. In virtual reality, the classical tasks that require such kinds of user command were early categorized in four [49]: navigating the virtual world, selecting a virtual object, manipulating it, or controlling the application (entering text, activating options, etc). The addition of a third dimension, the use of stereoscopic rendering and the use of advanced VR interfaces make however inappropriate many techniques that proved efficient in 2D, and make it necessary to design specific interaction techniques and adapted tools. This challenge is here renewed by the various kinds of 3D interaction which are targeted. In our case, we consider various cases, with motor and/or cerebral inputs, and potentially multiple users.
- **Modeling and simulation of complex 3D scenarios:** This second step corresponds to the update of the state of the virtual environment, in real-time, in response to all the potential commands or actions sent by the user. The complexity of the data and phenomena involved in 3D scenarios is constantly increasing. It corresponds for instance to the multiple states of the entities present in the simulation (rigid, articulated, deformable, fluids, which can constitute both the user's virtual body and the different manipulated objects), and the multiple physical phenomena implied by natural human interactions (squeezing, breaking, melting, etc). The challenge consists here in modeling and simulating these complex 3D scenarios and meeting, at the same time, two strong constraints of virtual reality systems: performance (real-time and interactivity) and genericity (e.g., multi-resolution, multi-modal, multi-platform, etc).
- **Immersive sensory feedbacks:** This third step corresponds to the display of the multiple sensory feedbacks (output) coming from the various VR interfaces. These feedbacks enable the user to perceive the changes occurring in the virtual environment. They are closing the 3D interaction loop, making the user immersed, and potentially generating a subsequent feeling of presence. Among the various VR interfaces which have been developed so far we can stress two kinds of sensory feedback: visual feedback (3D stereoscopic images using projection-based systems such as CAVE systems or Head Mounted Displays); and haptic feedback (related to the sense of touch and to tactile or force-feedback devices). The Hybrid team has a strong expertize in haptic feedback, and in the design of haptic and "pseudo-haptic" rendering [50]. Note that a major trend in the community, which is strongly supported by the Hybrid team, relates to a "perception-based" approach, which aims at designing sensory feedbacks which are well in line with human perceptual capacities.

These three scientific challenges are addressed differently according to the context and the user inputs involved. We propose to consider three different contexts, which correspond to the three different research axes of the Hybrid research team, namely : 1) body-based interaction (motor input only), 2) mind-based interaction (cerebral input only), and then 3) hybrid and collaborative interaction (i.e., the mixing of body and brain inputs from one or multiple users).

3.2. Research Axes

The scientific activity of Hybrid team follows three main axes of research:

- **Body-based interaction in virtual reality.** Our first research axis concerns the design of immersive and effective "body-based" 3D interactions, i.e., relying on a physical engagement of the user's body. This trend is probably the most popular one in VR research at the moment. Most VR setups make use of tracking systems which measure specific positions or actions of the user in order to interact with a virtual environment. However, in recent years, novel options have emerged for measuring

"full-body" movements or other, even less conventional, inputs (e.g. body equilibrium). In this first research axis we are thus concerned by the emergence of new kinds of "body-based interaction" with virtual environments. This implies the design of novel 3D user interfaces and novel 3D interactive techniques, novel simulation models and techniques, and novel sensory feedbacks for body-based interaction with virtual worlds. It involves real-time physical simulation of complex interactive phenomena, and the design of corresponding haptic and pseudo-haptic feedback.

- **Mind-based interaction in virtual reality.** Our second research axis concerns the design of immersive and effective "mind-based" 3D interactions in Virtual Reality. Mind-based interaction with virtual environments is making use of Brain-Computer Interface technology. This technology corresponds to the direct use of brain signals to send "mental commands" to an automated system such as a robot, a prosthesis, or a virtual environment. BCI is a rapidly growing area of research and several impressive prototypes are already available. However, the emergence of such a novel user input is also calling for novel and dedicated 3D user interfaces. This implies to study the extension of the mental vocabulary available for 3D interaction with VE, then the design of specific 3D interaction techniques "driven by the mind" and, last, the design of immersive sensory feedbacks that could help improving the learning of brain control in VR.
- **Hybrid and collaborative 3D interaction.** Our third research axis intends to study the combination of motor and mental inputs in VR, for one or multiple users. This concerns the design of mixed systems, with potentially collaborative scenarios involving multiple users, and thus, multiple bodies and multiple brains sharing the same VE. This research axis therefore involves two interdependent topics: 1) collaborative virtual environments, and 2) hybrid interaction. It should end up with collaborative virtual environments with multiple users, and shared systems with body and mind inputs.

IMAGINE Project-Team

3. Research Program

3.1. Methodology

As already stressed, thinking of future digital modeling technologies as an Expressive Virtual Pen enabling to seamlessly design, refine and convey animated 3D content, leads to revisit models for shapes, motions and stories from a user-centered perspective. More specifically, inspiring from the user-centered interfaces developed in the Human Computer Interaction domain, we introduced the new concept of user-centered graphical models. Ideally, such models should be designed to behave, under any user action, the way a human user would have predicted. In our case, user's actions may include creation gestures such as sketching to draft a shape or direct a motion, deformation gestures such as stretching a shape in space or a motion in time, or copy-paste gestures to transfer some of the features from existing models to other ones. User-centered graphical models need to incorporate knowledge in order to seamlessly generate the appropriate content from such actions. We are using the following methodology to advance towards these goals:

- Develop high-level models for shapes, motion and stories that embed the necessary knowledge to respond as expected to user actions. These models should provide the appropriate handles for conveying the user's intent while embedding procedural methods that seamlessly take care of the appropriate details and constraints.
- Combine these models with expressive design and control tools such as gesture-based control through sketching, sculpting, or acting, towards interactive environments where users can create a new virtual scene, play with it, edit or refine it, and semi-automatically convey it through a video.

3.2. Validation

Validation is a major challenge when developing digital creation tools: there is no ideal result to compare with, in contrast with more standard problems such as reconstructing existing shapes or motions. Therefore, we had to think ahead about our validation strategy: new models for geometry or animation can be validated, as usually done in Computer Graphics, by showing that they solve a problem never tackled before or that they provide a more general or more efficient solution than previous methods. The interaction methods we are developing for content creation and editing rely as much as possible on existing interaction design principles already validated withing the HCI community. We also occasionally develop new interaction tools, most often in collaboration with this community, and validate them through user studies. Lastly, we work with expert users from various application domains through our collaborations with professional artists, scientists from other domains, and industrial partners: these expert users validate the use of our new tools compared to their usual pipeline.

IN-SITU Project-Team

3. Research Program

3.1. Multi-disciplinary Research

InSitu uses a multi-disciplinary research approach, including computer scientists, psychologists and designers. Working together requires an understanding of each other's methods. Much of computer science relies on formal theory, which, like mathematics, is evaluated with respect to its internal consistency. The social sciences are based more on descriptive theory, attempting to explain observed behaviour, without necessarily being able to predict it. The natural sciences seek predictive theory, using quantitative laws and models to not only explain, but also to anticipate and control naturally occurring phenomena. Finally, design is based on a corpus of accumulated knowledge, which is captured in design practice rather than scientific facts but is nevertheless very effective.

Combining these approaches is a major challenge. We are exploring an integrative approach that we call *generative theory*, which builds upon existing knowledge in order to create new categories of artefacts and explore their characteristics Our goal is to produce prototypes, research methods and software tools that facilitate the design, development and evaluation of interactive systems [34].

LAGADIC Project-Team

3. Research Program

3.1. Visual servoing

Basically, visual servoing techniques consist in using the data provided by one or several cameras in order to control the motions of a dynamic system [1]. Such systems are usually robot arms, or mobile robots, but can also be virtual robots, or even a virtual camera. A large variety of positioning tasks, or mobile target tracking, can be implemented by controlling from one to all the degrees of freedom of the system. Whatever the sensor configuration, which can vary from one on-board camera on the robot end-effector to several free-standing cameras, a set of visual features has to be selected at best from the image measurements available, allowing to control the desired degrees of freedom. A control law has also to be designed so that these visual features $\mathbf{s}(t)$ reach a desired value $\mathbf{s}^*$, defining a correct realization of the task. A desired planned trajectory $\mathbf{s}^*(t)$ can also be tracked. The control principle is thus to regulate to zero the error vector $\mathbf{s}(t) - \mathbf{s}^*(t)$. With a vision sensor providing 2D measurements, potential visual features are numerous, since 2D data (coordinates of feature points in the image, moments, ...) as well as 3D data provided by a localization algorithm exploiting the extracted 2D features can be considered. It is also possible to combine 2D and 3D visual features to take the advantages of each approach while avoiding their respective drawbacks.

More precisely, a set $\mathbf{s}$ of k visual features can be taken into account in a visual servoing scheme if it can be written:

$$\mathbf{s} = \mathbf{s}(\mathbf{x}(\mathbf{p}(t)), \mathbf{a}) \tag{91}$$

where $\mathbf{p}(t)$ describes the pose at the instant t between the camera frame and the target frame, $\mathbf{x}$ the image measurements, and $\mathbf{a}$ a set of parameters encoding a potential additional knowledge, if available (such as for instance a coarse approximation of the camera calibration parameters, or the 3D model of the target in some cases).

The time variation of $\mathbf{s}$ can be linked to the relative instantaneous velocity $\mathbf{v}$ between the camera and the scene:

$$\dot{\mathbf{s}} = \frac{\partial \mathbf{s}}{\partial \mathbf{p}} \dot{\mathbf{p}} = \mathbf{L}_{\mathbf{s}} \mathbf{v} \tag{92}$$

where $\mathbf{L}_{\mathbf{s}}$ is the interaction matrix related to $\mathbf{s}$. This interaction matrix plays an essential role. Indeed, if we consider for instance an eye-in-hand system and the camera velocity as input of the robot controller, we obtain when the control law is designed to try to obtain an exponential decoupled decrease of the error:

$$\mathbf{v}_c = -\lambda \widehat{\mathbf{L}_{\mathbf{s}}}^{+} (\mathbf{s} - \mathbf{s}^*) - \widehat{\mathbf{L}_{\mathbf{s}}}^{+} \widehat{\frac{\partial \mathbf{s}}{\partial t}} \tag{93}$$

where λ is a proportional gain that has to be tuned to minimize the time-to-convergence, $\widehat{\mathbf{L}_{\mathbf{s}}}^{+}$ is the pseudo-inverse of a model or an approximation of the interaction matrix, and $\widehat{\frac{\partial \mathbf{s}}{\partial t}}$ an estimation of the features velocity due to a possible own object motion.

From the selected visual features and the corresponding interaction matrix, the behavior of the system will have particular properties as for stability, robustness with respect to noise or to calibration errors, robot 3D trajectory, etc. Usually, the interaction matrix is composed of highly non linear terms and does not present any decoupling properties. This is generally the case when $\mathbf{s}$ is directly chosen as $\mathbf{x}$. In some cases, it may lead to inadequate robot trajectories or even motions impossible to realize, local minimum, tasks singularities, etc. It is thus extremely important to design adequate visual features for each robot task or application, the ideal case (very difficult to obtain) being when the corresponding interaction matrix is constant, leading to a simple linear control system. To conclude in few words, **visual servoing is basically a non linear control problem. Our Holy Grail quest is to transform it into a linear control problem.**

Furthermore, embedding visual servoing in the task function approach allows solving efficiently the redundancy problems that appear when the visual task does not constrain all the degrees of freedom of the system. It is then possible to realize simultaneously the visual task and secondary tasks such as visual inspection, or joint limits or singularities avoidance. This formalism can also be used for tasks sequencing purposes in order to deal with high level complex applications.

3.2. Visual tracking

Elaboration of object tracking algorithms in image sequences is an important issue for researches and applications related to visual servoing and more generally for robot vision. A robust extraction and real time spatio-temporal tracking process of visual cues is indeed one of the keys to success of a visual servoing task. If fiducial markers may still be useful to validate theoretical aspects in modeling and control, natural scenes with non cooperative objects and subject to various illumination conditions have to be considered for addressing large scale realistic applications.

Most of the available tracking methods can be divided into two main classes: feature-based and model-based. The former approach focuses on tracking 2D features such as geometrical primitives (points, segments, circles,...), object contours, regions of interest...The latter explicitly uses a model of the tracked objects. This can be either a 3D model or a 2D template of the object. This second class of methods usually provides a more robust solution. Indeed, the main advantage of the model-based methods is that the knowledge about the scene allows improving tracking robustness and performance, by being able to predict hidden movements of the object, detect partial occlusions and acts to reduce the effects of outliers. The challenge is to build algorithms that are fast and robust enough to meet our applications requirements. Therefore, even if we still consider 2D features tracking in some cases, our researches mainly focus on real-time 3D model-based tracking, since these approaches are very accurate, robust, and well adapted to any class of visual servoing schemes. Furthermore, they also meet the requirements of other classes of application, such as augmented reality.

3.3. Slam

Most of the applications involving mobile robotic systems (ground vehicles, aerial robots, automated submarines,...) require a reliable localization of the robot in its environment. A challenging problem is when neither the robot localization nor the map is known. Localization and mapping must then be considered concurrently. This problem is known as Simultaneous Localization And Mapping (Slam). In this case, the robot moves from an unknown location in an unknown environment and proceeds to incrementally build up a navigation map of the environment, while simultaneously using this map to update its estimated position.

Nevertheless, solving the Slam problem is not sufficient for guaranteeing an autonomous and safe navigation. The choice of the representation of the map is, of course, essential. The representation has to support the different levels of the navigation process: motion planning, motion execution and collision avoidance and, at the global level, the definition of an optimal strategy of displacement. The original formulation of the Slam problem is purely metric (since it basically consists in estimating the Cartesian situations of the robot and a set of landmarks), and it does not involve complex representations of the environment. However, it is now well recognized that **several complementary representations are needed to perform exploration, navigation, mapping, and control tasks successfully. We propose to use composite models of the environment that**

mix topological, metric, and grid-based representations. Each type of representation is well adapted to a particular aspect of autonomous navigation: the metric model allows one to locate the robot precisely and plan Cartesian paths, the topological model captures the accessibility of different sites in the environment and allows a coarse localization, and finally the grid representation is useful to characterize the free space and design potential functions used for reactive obstacle avoidance. However, ensuring the consistency of these various representations during the robot exploration, and merging observations acquired from different viewpoints by several cooperative robots, are difficult problems. This is particularly true when different sensing modalities are involved. New studies to derive efficient algorithms for manipulating the hybrid representations (merging, updating, filtering...) while preserving their consistency are needed.

LEAR Project-Team

3. Research Program

3.1. Image features and descriptors and robust correspondence

Reliable image features are a crucial component of any visual recognition system. Despite much progress, research is still needed in this area. Elementary features and descriptors suffice for a few applications, but their lack of robustness and invariance puts a heavy burden on the learning method and the training data, ultimately limiting the performance that can be achieved. More sophisticated descriptors allow better inter-class separation and hence simpler learning methods, potentially enabling generalization from just a few examples and avoiding the need for large, carefully engineered training databases.

The feature and descriptor families that we advocate typically share several basic properties:

- **Locality and redundancy:** For resistance to variable intra-class geometry, occlusions, changes of viewpoint and background, and individual feature extraction failures, descriptors should have relatively small spatial support and there should be many of them in each image. Schemes based on collections of image patches or fragments are more robust and better adapted to object-level queries than global whole-image descriptors. A typical scheme thus selects an appropriate set of image fragments, calculates robust appearance descriptors over each of these, and uses the resulting collection of descriptors as a characterization of the image or object (a "bag-of-features" approach – see below).
- **Photometric and geometric invariance:** Features and descriptors must be sufficiently invariant to changes of illumination and image quantization and to variations of local image geometry induced by changes of viewpoint, viewing distance, image sampling and by local intra-class variability. In practice, for local features geometric invariance is usually approximated by invariance to Euclidean, similarity or affine transforms of the local image.
- **Repeatability and salience:** Fragments are not very useful unless they can be extracted reliably and found again in other images. Rather than using dense sets of fragments, we often focus on local descriptors based at particularly salient points – "keypoints" or "points of interest". This gives a sparser and thus potentially more efficient representation, and one that can be constructed automatically in a preprocessing step. To be useful, such points must be accurately relocalizable in other images, with respect to both position and scale.
- **Informativeness:** Notwithstanding the above forms of robustness, descriptors must also be informative in the sense that they are rich sources of information about image content that can easily be exploited in scene characterization and object recognition tasks. Images contain a lot of variety so high-dimensional descriptions are required. The useful information should also be manifest, not hidden in fine details or obscure high-order correlations. In particular, image formation is essentially a spatial process, so relative position information needs to be made explicit, e.g. using local feature or context style descriptors.

Partly owing to our own investigations, features and descriptors with some or all of these properties have become popular choices for visual correspondence and recognition, particularly when large changes of viewpoint may occur. One notable success to which we contributed is the rise of "bag-of-features" methods for visual object recognition. These characterize images by their (suitably quantized or parametrized) global distributions of local descriptors in descriptor space. The representation evolved from texton based methods in texture analysis. Despite the fact that it does not (explicitly) encode much spatial structure, it turns out to be surprisingly powerful for recognizing more structural object categories.

Our current research on local features is focused on creating detectors and descriptors that are better adapted to describe object classes, on incorporating spatial neighborhood and region constraints to improve informativeness relative to the bag-of-features approach, and on extending the scheme to cover different kinds of locality. Current research also includes the development and evaluation of local descriptors for video, and associated detectors for spatio-temporal content.

3.2. Statistical modeling and machine learning for image analysis

We are interested in learning and statistics mainly as technologies for attacking difficult vision problems, so we take an eclectic approach, using a broad spectrum of techniques ranging from classical statistical generative and discriminative models to modern kernel, margin and boosting based approaches. Hereafter we enumerate a set of approaches that address some problems encountered in this context.

- Parameter-rich models and limited training data are the norm in vision, so overfitting needs to be estimated by cross-validation, information criteria or capacity bounds and controlled by regularization, model and feature selection.
- Visual descriptors tend to be high-dimensional and redundant, so we often preprocess data to reduce it to more manageable terms using dimensionality reduction techniques including PCA and its non-linear variants, latent structure methods such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA), and manifold methods such as Isomap/LLE.
- To capture the shapes of complex probability distributions over high-dimensional descriptor spaces, we either fit mixture models and similar structured semi-parametric probability models, or reduce them to histograms using vector quantization techniques such as K-means or latent semantic structure models.
- Missing data is common owing to unknown class labels, feature detection failures, occlusions and intra-class variability, so we need to use data completion techniques based on variational methods, belief propagation or MCMC sampling.
- Weakly labeled data is also common – for example one may be told that a training image contains an object of some class, but not where the object is in the image – and variants of unsupervised, semi-supervised and co-training are useful for handling this. In general, it is expensive and tedious to label large numbers of training images so less supervised data mining style methods are an area that needs to be developed.
- On the discriminative side, machine learning techniques such as Support Vector Machines, Relevance Vector Machines, and Boosting, are used to produce flexible classifiers and regression methods based on visual descriptors.
- Visual categories have a rich nested structure, so techniques that handle large numbers of classes and nested classes are especially interesting to us.
- Images and videos contain huge amounts of data, so we need to use algorithms suited to large-scale learning problems.

3.3. Visual recognition and content analysis

Current progress in visual recognition shows that combining advanced image descriptors with modern learning and statistical modeling techniques is producing significant advances. We believe that, taken together and tightly integrated, these techniques have the potential to make visual recognition a mainstream technology that is regularly used in applications ranging from visual navigation through image and video databases to human-computer interfaces and smart rooms.

The recognition strategies that we advocate make full use of the robustness of our invariant image features and the richness of the corresponding descriptors to provide a vocabulary of base features that already goes a long way towards characterizing the category being recognized. Trying to learn everything from scratch using simpler, non-invariant features would require far too much data: good learning cannot easily make up for bad features. The final classifier is thus responsible “only” for extending the base results to larger amounts of intra-class and viewpoint variation and for capturing higher-order correlations that are needed to fine tune the performance.

That said, learning is not restricted to the classifier and feature sets can not be designed in isolation. We advocate an end-to-end engineering approach in which each stage of the processing chain combines learning with well-informed design and exploitation of statistical and structural domain models. Each stage is thoroughly tested to quantify and optimize its performance, thus generating or selecting robust and informative features, descriptors and comparison metrics, squeezing out redundancy and bringing out informativeness.

LINKMEDIA Project-Team

3. Research Program

3.1. Scientific background

LINKMEDIA is a multidisciplinary research team, with multimedia data as the main object of study. We are guided by the data and their specificity—semantically interpretable, heterogeneous and multimodal, available in large amounts, unstructured and disconnected—, as well as by the related problems and applications.

With multimedia data at the center, orienting our choices of methods and algorithms and serving as a basis for experimental validation, the team is directly contributing to the following scientific fields:

- multimedia: content-based analysis; multimodal processing and fusion; multimedia applications;
- computer vision: compact description of images; object and event detection;
- natural language processing: topic segmentation; information extraction;
- information retrieval: high-dimensional indexing; approximate k-nn search; efficient set comparison;

LINKMEDIA also takes advantage of advances in the following fields, adapting recent developments to the multimedia area:

- signal processing – image processing; compression;
- machine learning – deep architectures; structured learning; adversarial learning;
- security – data encryption; differential privacy;
- data mining – time series mining and alignment; pattern discovery; knowledge extraction;

3.2. Workplan

Research activities in LINKMEDIA are organized along three major lines of research which build upon the scientific domains already mentioned.

3.2.1. Unsupervised motif discovery

As an alternative to supervised learning techniques, unsupervised approaches have emerged recently with the goal of discovering directly patterns and events of interest from the data, in a totally unsupervised manner. In the absence of prior knowledge on what we are interested in, meaningfulness can be judged based on one of three main criteria: unexpectedness, saliency and recurrence. This last case posits that repeating patterns, known as motifs, are potentially meaningful, leading to recent work on the unsupervised discovery of motifs in multimedia data [77], [75], [76].

LINKMEDIA seeks to *develop unsupervised motif discovery approaches which are both accurate and scalable*. In particular, we consider the discovery of repeating objects in image collections and the discovery of repeated sequences in video and audio streams. Research activities are organized along the following lines:

- developing the scientific basis for scalable motif discovery: sparse histogram representations; efficient co-occurrence counting; geometry and time aware indexing schemes;
- designing and evaluating accurate and scalable motif discovery algorithms applied to a variety of multimedia content: exploiting efficient geometry or time aware matching functions; fast approximate DTW; symbolic representations of multimedia data, in conjunction with existing symbolic data mining approaches;
- developing methodology for the interpretation, exploitation and evaluation of motif discovery algorithms in various use-cases: image classification; video stream monitoring; transcript-free NLP for spoken document;

3.2.2. Describing and structuring

Content-based analysis has received a lot of attention from the early days of multimedia, with an extensive use of supervised machine learning for all modalities [78], [72]. Progress in large scale entity and event recognition in multimedia content has made available general purpose approaches able to learn from very large data sets and performing fairly decently in a large number of cases. Current solutions are however limited to simple, homogeneous, information and can hardly handle structured information such as hierarchical descriptions, tree-structured or nested concepts.

LINKMEDIA aims at *expanding techniques for multimedia content modeling, event detection and structure analysis*. The main transverse research lines that LINKMEDIA will develop are as follows:

- context-aware content description targeting (homogeneous) collections of multimedia data: latent variable discovery; deep feature learning; motif discovery;
- secure description to enable privacy and security aware multimedia content processing: everaging encryption and diversity; exploring adversarial machine learning in a multimedia context; privacy-oriented image processing;
- multilevel modeling with a focus on probabilistic modeling of structured multimodal data: multiple kernels; structured machine learning; conditionnal random fields;

3.2.3. Linking

Creating explicit links between media content items has been considered on different occasions, with the goal of seeking and discovering information by browsing, as opposed to information retrieval via ranked lists of relevant documents. Content-based link creation has been initially addressed in the hypertext community for well-structured texts [71] and was recently extended to multimedia content [79], [74], [73]. The problem of organizing collections with links remains mainly unsolved for large heterogeneous collections of unstructured documents, with many issues deserving attention: linking at a fine semantic grain; selecting relevant links; characterizing links; evaluating links; etc.

LINKMEDIA targets pioneering research on media linking by **developing scientific ground, methodology and technology for content-based media linking** directed to applications exploiting rich linked content such as navigation or recommendation. Contributions are concentrated along the following lines:

- algorithmic of linked media for content-based link authoring in multimedia collections: time-aware graph construction; multimodal hypergraphs; large scale k-nn graphs;
- link interpretation and characterization to provide links semantics for interpretability: text alignment; entity linking; intention vs. extension;
- linked media usage and evaluation: information retrieval; summarization; data models for navigation; link prediction;

LINKS Team (section vide)

MAGNET Team

3. Research Program

3.1. Introduction

The main objective of MAGNET is to develop original machine learning methods for networked data. We consider information networks in which the data are vectorial data and texts. We model such information networks as (multiple) (hyper)graphs wherein nodes correspond to entities (documents, spans of text, users, ...) and edges correspond to relations between entities (similarity, answer, co-authoring, friendship, ...). Our main research goal is to propose new learning algorithms to build applications like browsing, monitoring and recommender systems, and more broadly information extraction in information networks. Hence, we will investigate new learning algorithms for node clustering and node classification, link classification and link prediction. Also, we will search for the best hidden graph structure to be generated for solving a given learning task. We will base our research on generative models for graphs, on machine learning for graphs and on machine learning for texts. The challenges are the dimensionality of the input space, possibly the dimensionality of the output space, the high level of dependencies between the data, the inherent ambiguity of textual data and the limited amount of human labeling. An additional challenge will be to design scalable methods for large information networks. Hence, we will explore how sampling and randomization can be used in new machine learning algorithms. Also, active machine learning algorithms for graphs will be investigated.

On the first hand we want to design machine learning algorithms on graphs to solve problems in networks of texts and documents in natural language. The main originality of this research is to consider and take advantage of the setting of networked data exploiting the relationships between different data entities and, overall, the graph topology. On the second hand, in a concomitant way, we want to develop prediction models for graph-like data. This includes prediction, ranking and classification of links and nodes in an on-line or batch setting. The two objectives are intertwined, enrich each other and raise important scientific questions we want to focus on. Our research proposal is organized according to the following questions:

1. How to go beyond vectorial classification models in natural language oriented tasks?
2. How to adaptively build graphs with respect to the given tasks? How to create network from observations of information diffusion processes?
3. How to design methods able to achieve very good predictive accuracy without giving up on scalability?
4. How to go beyond strict node homophilic/similarity assumptions in graph-based learning methods?

3.2. Beyond vectorial models for NLP

One of our overall research objectives is to derive graph-based machine learning algorithms for natural language and text information extraction tasks. This section discusses the motivations behind the use of graph-based ML approaches for these tasks, the main challenges associated with it, as well as some concrete projects. Some of the challenges go beyond NLP problems and will be further developed in the next sections. An interesting aspect of the project is that we anticipate some important cross-fertilizations between NLP and ML graph-based techniques, with NLP not only benefiting from but also pushing ML graph-based approaches into new directions.

Motivations for resorting to graph-based algorithms for texts are at least threefold. First, online texts are organized in networks. With the advent of the web, and the development of forums, blogs, and micro-blogging, and other forms of social media, text productions have become strongly connected. Thus, documents on the web are linked through hyperlinks, forum posts and emails are organized in threads, tweets can be retweeted, etc. Additional connections can be made through users connections (co-authorship, friendship, follower, etc.). Interestingly, NLP research has been rather slow in coming to terms with this situation, and most work still focus on document-based or sentence-based predictions (wherein inter-document or inter-sentence structure is not exploited). Furthermore, several multi-document tasks exist in NLP (such as multi-document summarization and cross-document coreference resolution), but most existing work typically ignore document boundaries and simply apply a document-based approach, therefore failing to take advantage of the multi-document dimension [26], [28].

A second motivation comes from the fact that most (if not all) NLP problems can be naturally conceived as graph problems. Thus, NL tasks often involve discovering a relational structure over a set of text spans (words, phrases, clauses, sentences, etc.). Furthermore, the *input* of numerous NLP tasks is also a graph; indeed, most end-to-end NLP systems are conceived as pipelines wherein the output of one processor is in the input of the next. For instance, several tasks take POS tagged sequences or dependency trees as input. But this structured input is often converted to a vectorial form, which inevitably involves a loss of information.

Finally, graph-based representations and learning methods in principle appear to address some core problems faced by NLP, such as the fact that textual data are typically not independent and identically distributed, they often live on a manifold, they involve very high dimensionality, and their annotations is costly and scarce. As such, graph-based methods represent an interesting alternative, or at least complement, to structured prediction methods (such as CRFs or structured SVMs) commonly used within NLP. While structured output approaches are able to model local dependencies (e.g., between neighboring words or sentences), they cannot efficiently capture long distance dependencies, like forcing a particular n-gram to receive the same labeling in different sentences or documents for instance. On the other hand, graph-based models provide a natural way to capture global properties of the data through the exploitation of walks and neighborhood in graphs. Graph-based methods, like label propagation, have also been shown to be very effective in semi-supervised settings, and have already given some positive results on a few NLP tasks [9], [30].

Given the above motivations, our first line of research will be to investigate how one can leverage an underlying network structure (e.g., hyperlinks, user links) between documents, or text spans in general, to enhance prediction performances for several NL tasks. We think that a "network effect", similar to the one that took place in Information Retrieval (with the Page Rank algorithm), could also positively impact NLP research. A few recent papers have already opened the way, for instance in attempting to exploit Twitter follower graph to improve sentiment classification [29].

Part of the challenge in this work will be to investigate how adequately and efficiently one can model these problems as instances of more general graph-based problems, such as node clustering/classification or link prediction discussed in the next sections. In a few cases, like text classification or sentiment analysis, graph modeling appears to be straightforward: nodes correspond to texts (and potentially users), and edges are given by relationships like hyperlinks, co-authorship, frienship, or thread membership. Unfortunately, modeling NL problems as networks is not always that obvious. From the one hand, the right level of representation will probably vary depending on the task at hand: the nodes will be sentences, phrases, words, etc. From the other hand, the underlying graph will typically not be given a priori, which in turn raises the question of how we construct it. Of course, there are various well-known ways to obtain similarity measures between text contents (and its associated vectorial data), and graphs can be easily constructed from those combined with some sparsification method. But we would like our similarity to be tailored to the task objective. An additional problem with many NLP problems is that features typically live in different types of spaces (e.g., binary, discrete, continuous). A preliminary discussion of the issue of optimal graph construction for semi-supervised learning in NLP is given in [9], [33]. We identify the issue of adaptative graph construction as an important scientific challenge for machine learning on graphs in general, and we will discuss it further in Section 3.3 .

As noted above, many NLP tasks have been recast as structure prediction problems, allowing to capture (some of the) output dependencies. Structure prediction can be viewed as (set of) link prediction with global loss or dependencies, which means that graph-based learning methods can handle (at least, approximately) output prediction dependencies, and they can in principle capture additional more global dependencies given the right graph structure. How to best combine structured output and graph-based ML approaches is another challenge that we intend to address. We will initially investigate this question within a semi-supervised context, concentrating on graph based regularization and graph propagation methods. Within such approaches, labels are typically binary or they correspond to small finite set. Our objective is to explore how one propagates an exponential number of *structured labels* (like a sequence of tags or a dependency tree) through graphs. Recent attempts at blending structured output models with graph-based models are investigated in [30], [17]. Another related question that we will address in this context is how does one learn with *partial labels* (like partially specificied tag sequence or tree) and use the graph structure to complete the output structure. This last question is very relevant to NL problems where human annotations are costly; being able to learn from partial annotations could therefore allow for more targeted annotations and in turn reduced costs [18].

The NL tasks we will mostly focus on are coreference resolution and entity linking, temporal structure prediction, and discourse parsing. These tasks will be envisonned in both document and cross-document settings, although we expect to exploit inter-document links either way. Choices for these particular tasks is guided by the fact that are still open problems for the NLP community, they potentially have a high impact for industrial applications (like information retrieval, question answering, etc.), and we already have some expertise on these tasks in the team. As a midterm goal, we also plan to work on tasks more directly relating to micro-blogging, such sentiment analysis and the automatic thread structuring of technical forums; the latter task is in fact an instance of rhetorical structure prediction [32].

We have already initiated some work on the coreference resolution problem in the context of ML graph-based approaches. We cast this problem as a spectral clustering problem. Given than features can be numerical or nominal, the definition of a good similarity measure between entities is not straightforward. As a first solution, we consider only numerical attributes to build a k-nn graph of mentions so that graph clustering methods can be applied. Nominal attributes and relations are introduced by means of soft constraints on this clustering. Constraints can have various forms and have the ability of going beyond homophily assumptions, taking into account for instance dissimilarity relationships. From this setting we derive new graph-based learning methods. We propose to study the modification of graph clustering and spectral embeddings to satisfy certain constraints induced by several types of supervision: (i) nodes belong to the same group or to different groups, and (ii) some groups are fully known while others have to be discovered. This semi-supervised graph clustering problem is studied in a batch and transductive setting. But interesting extensions can be investigated in an online and active setting.

3.3. Adaptive Graph Construction

In most applications, edge weights are computed through a complex data-modeling process and convey crucially important information for classifying nodes, which makes it possible to infer information related to each data sample even exploiting the graph topology solely. In fact, a widespread approach to the solution of several classification problems is representing the data through an undirected weighted graph in which edge weights quantify the similarity between data points. This technique for coding input data has been applied to several domains, including classification of genomic data ([27]), face recognition ([16]), and text categorization ([21]).

In some cases, the full adjacency matrix is generated by employing suitable similarity functions chosen through a deep understanding of the problem structure. For example TF-IDF representation of documents, the affinity between pairs of samples is often estimated through the cosine measure or the χ^2 distance. After the generation of the full adjacency matrix, the second phase for obtaining the final graph consists in an edge sparsification/reweighting operation. Some of the edges of the clique obtained in the first step are pruned and the remaining ones can be reweighted to meet the specific requirements of the given classification problem. Constructing a graph with these methods obviously entails various kinds of loss of information. However, in

problems like node classification, the use of graphs generated from several datasets can lead to an improvement in accuracy performance ([34], [10], [11]). Hence, the transformation of a dataset into a graph may, at least in some cases, partially remove various kinds of irregularities present in the original datasets, while keeping some of the most useful information for classifying the data samples. Moreover, it is often possible to accomplish classification tasks on the obtained graph using a running time remarkably lower than is needed by algorithms exploiting the initial datasets, and a suitable sparse graph representation can be seen as a compressed version of the original data. This holds even when input data are provided in a online/stream fashion, so that the resulting graph evolves over time.

In this project we will address the problem of adaptive graph construction towards several directions. One is the question of choosing the best similarity measure given the objective learning task. This question is related to the question of similarity learning ([12]) which has not been considered in the context of graph based learning. In the context of structured prediction, we will develop approaches where output structures are organized in graphs whose similarity is given by top-k outcomes of greedy algorithms.

A different way we envision adaptative graph construction is in the context of semi-supervised learning. Partial supervision can take various forms and an interesting and original setting is governed by two currently studied applications: detection of brain anomaly from connectome data and polls recommendation in marketing. Indeed, for these two applications, a partial knowledge of the information diffusion process can be observed while the network is unknown or only partially known. An objective is to construct (or complete) the network structure from some local diffusion information. The problem can be formalized as a graph construction problem from partially observed diffusion processes. It has been studied very recently in [23]. In our case, the originality comes either from the existence of different sources of observations or from the large impact of node contents in the network.

We will study how to combine graphs defined by networked data and graphs built from flat data to solve a given task. This is of major importance for information networks because, as said above, we will have to deal with multiple relations between entities (texts, spans of texts, ...) and also use textual data and vectorial data. We have started to work on combining graphs in a semi supervised setting for node classification problems along the PhD thesis of T. Ricatte. Future work include combination geared by semi-supervision on link prediction tasks. This can be studied in an active learning setting. But one important issue is to design scalable approaches, thus to exploit locality given by the network. Doing this we address another objective to build non uniformly parameterized combinations.

3.4. Prediction on Graphs and Scalability

As stated in the previous sections, graphs as complex objects provides a rich representation of data. Often enough the data is only partially available and the graph representation is very helpful in predicting the unobserved elements. We are interested in problems where the complete structure of the graph needs to be recover and only a fraction of the links is observed. The link prediction problem falls into this category. We are also interested in the recommendation and link classification problems which can be seen as graphs where the structure is complete but some labels on the links (weights or signs) are missing. Finally we are also interested in labelling the nodes of the graph, with class or cluster memberships or with a real value, provided that we have (some information about) the labels for some of the nodes.

The semi-supervised framework will be also considered. A midterm research plan is to study how graph-based regularization models help for structured prediction problems. This question will be studied in the context of NLP tasks, as noted in Section 3.2 , but we also plan to develop original machine learning algorithms that have a more general applicability. Inputs are networks whose nodes (texts) have to be labeled by structures. We assume that structures lie in some manifold and we want to study how labels can propagate in the network. One approach is to find smooth labeling function corresponding to an harmonic function on both manifolds in input and output. We also plan to extend our results on spectral clustering with must-link and cannot-link constraints in two directions. We have proposed a batch method with an optimization problem based on an adaptive spectral embedding with respects to constraints. We want to extend this approach to an on-line and

active setting where a flow of graphs (each one is a document) is given as input. In the case of large graphs, we also consider the case where partial supervision consists in the knowledge of few clusters.

Scalability is one of the main issue in the design of new prediction algorithms working on networked data. It has gained more and more importance in recent years, because of the growing size of the most popular networked data that are now used by millions of people. In such contexts, learning algorithms whose computation time scales quadratically, or slower, in the number of considered data objects (usually nodes or vertices, depending on the given task) should be considered impractical.

These observations lead to the idea of using graph sparsification techniques in order to work on a part of the original network for getting results that can be easily extended and used for the whole original input. A sparsified version of the original graph can often be seen as a subset of the initial input, i.e. a suitably selected input subgraph which forms the training set (or, more in general, it is included in the training set). This holds even for the active setting.

A simple example could be to find a spanning tree of the input graph, possibly using randomization techniques, with properties such that we are allowed to obtain interesting results for the initial graph dataset. We have started to explore this research direction for instance in [31]. This approach leaves us with the problem of choosing a good spanning tree, taking into account that the setting could be adversarial (e.g, in the online case the presentation and the assignment of the labels are both arbitrary). A suitable use of the randomization power becomes therefore remarkably significant. Moreover, it is interesting to observe that running a prediction algorithm on a sparsified version of the input dataset allows the parallelization of prediction tasks. In fact, given a prediction task for a networked dataset, in a preliminary phase one could run a randomized graph sparsification method in parallel on different machines. For example, in the case of the spanning tree use, one could then draw several spanning trees at the same time, each on a different computer. This way it is possible to simultaneously run different prediction experiments on the same task and aggregating the obtained results at the end, with several methods (e.g. simply by majority vote) in order to increase the robustness and accuracy predictions.

At the level of the mathematical foundations, the key issue to be addressed in the study of (large-scale) random networks also concerns the segmentation of network data into sets of independent and identically distributed observations. If we identify the data sample with the whole network, as it has been done in previous approaches [22], we typically end up with a set of observations (such as nodes or edges) which are highly interdependent and hence overly violate the classic i.i.d. assumption. In this case, the data scale can be so large and the range of correlations can be so wide, that the cost of taking into account the whole data and their dependencies is typically prohibitive. On the contrary, if we focus instead on a set of subgraphs independently drawn from a (virtually infinite) target network, we come up with a set of independent and identically distributed observations—namely the subgraphs themselves, where subgraph sampling is the underlying ergodic process [13]. Such an approach is one principled direction for giving novel statistical foundations to random network modeling. At the same time, because one shifts the focus from the whole network to a set of subgraphs, complexity issues can be restricted to the number of subgraphs and their size. The latter quantities can be controlled much more easily than the overall network size and dependence relationships, thus allowing to tackle scalability challenges through a radically redesigned approach.

We intend to develop new learning models for link prediction problems. We have already proposed a conditional model in [20] with statistics based on Fiedler values computed on small subgraphs. We will investigate the use of such a conditional model for link prediction. We will also extend the conditional probabilistic models to the case of graphs with textual and vectorial data by defining joint conditional models. Indeed, an important challenge for information networks is to introduce node contents in link ranking and link prediction methods that usually rely solely on the graph structure. A first step in this direction was already proposed in [19] where we learn a mapping of node content to a new representation constrained by the existing link structure and applied it for link recommendation. This approach opens a different view on recommendation by means of link ranking problems for which we think that non parametric approaches should be fruitful.

Regarding link classification problems, we plan to devise a whole family of active learning strategies, which could be based on spanning trees or sparse input subgraphs, that exploit randomization and the structure of the graph in order to offset the adversarial label assignment. We expect these active strategies to exhibit good accuracies with a remarkably small number of queried edges, where passive learning methods typically break down. The theoretical findings can be supported by experiments run on both synthetic and real-world (Slashdot, Epinions, Wikipedia, and others) datasets.

We are interested in studying generative models for graph labeling, exploiting the results obtained in p-stochastic model for link classification (investigated in [15]) and statistical model for node label assignment which can be related to tree-structured Markov random fields [24].

In developing our algorithms, we focus on providing theoretical guarantees on prediction accuracy and, at the same time, on computational efficiency. The development of methods that simultaneously guarantee optimal accuracy and computational efficiency is a very challenging goal. In fact, the accuracy of most methods in the literature is not rigorously analyzed from a theoretical point of view. Likewise, tight time and space complexity bounds are not generally provided. This contrasts with the need to manage extremely large relational datasets like, e.g., snapshots of the World Wide Web.

3.5. Beyond Homophilic Relationships

In many cases, the algorithms devised for solving node classification problems are driven by the following assumption: linked entities tend to be assigned to the same class. This assumption, in the context of social networks, is known as homophily ([14], [25]) and involves ties of every type, including friendship, work, marriage, age, gender, and so on. In social networks, homophily naturally implies that a set of individuals can be parted into subpopulations that are more cohesive. In fact, the presence of homogeneous groups sharing interests is one of the most significant reasons for affinity among interconnected individuals, which suggests that, in spite of its simplicity, this principle turns out to be very powerful for node classification problems in general networks.

Recently, however, researchers have started to consider networked data where connections may also carry a negative meaning. For instance, disapproval or distrust in social networks, negative endorsements on the Web. Concrete examples are provided by certain types of online social networks. Users of Slashdot can tag other users as friends or foes. Similarly, users of Epinions can give positive or negative ratings not only to products but also to other users. Even in the social network of Wikipedia administrators, votes cast by an admin in favor or against the promotion of another admin can be viewed as positive or negative links. More examples of signed links are found in other domains, such as the excitatory or inhibitory interactions between genes or gene products in biological networks.

Although the introduction of signs on graph edges appears like a small change from standard weighted graphs, the resulting mathematical object, called signed graph, has an unexpectedly rich additional complexity. For example, the spectral properties of signed graphs, which essentially all sophisticated node classification algorithms rely on, are different and less known than those of their unsigned counterparts. Signed graphs naturally lead to a specific inference problem that we have discussed in previous sections: link classification. This is the problem of predicting the sign of links in a given graph. In online social networks, this may be viewed as a form of sentiment analysis, since we would like to semantically categorize the relationship between individuals.

Another way to go beyond homophily between entities will be studied using our recent model of hypergraphs with bipartite hyperedges [4]. A bipartite hyperedge connects two ends which are disjoint subsets of nodes. Bipartite hyperedges is a way to relate two collections of (possibly heterogeneous) entities represented by nodes. In the NLP setting, while hyperedges can be used to model bags of words, bipartite hyperedges are associated with relationships between bags of words. But each end of bipartite hyperedges is also a way to represent complex entities, gathering several attribute values (nodes) into hyperedges viewed as records. Our hypergraph notion naturally extends directed and undirected weighted graph. We have defined a spectral theory for this new class of hypergraphs and opened a way to smooth labeling on sets of nodes. The weighting

scheme permits to weight the participation of each node to the relationship modeled by bipartite hyperedges accordingly to an equilibrium condition. This is exactly that equilibrium condition that provides a competition between nodes in hyperedges and allows interesting modeling properties that go beyond homophily and similarity over nodes. (Theoretical analysis of our hypergraphs exhibits tight relationships with signed graphs). Following this competition idea, bipartite hyperedges are like matches between two teams and examples of applications are team creation. The basic tasks in which we are interested in are hyperedge classification, hyperedge prediction, node weight prediction. Finally, hypergraphs also represent a way to summarize or compress large graphs in which there exists highly connected couples of (large) subsets of nodes.

To conclude, we plan to go beyond the homophilic bias from the algorithmic as well as from the modeling point of view. We will consider new kind of modeling and learning biases provided by graphs with negative weights (signed graphs) and hypergraphs. We will study their spectral properties, smoothness measures of (node or edge) labeling. Sampling and walking also need to be reconsidered. From the machine learning perspective, we will study edge and node labeling in batch and online settings. In connection with our main targeted applications, we will mainly consider unsupervised and semi-supervised situations. We think that allowing negative weights and advanced relationships on nodes will also lead to space efficient representations of graphs.

MAGRIT Project-Team

3. Research Program

3.1. Matching and 3D tracking

One of the most basic problems currently limiting AR applications is the registration problem. The objects in the real and virtual worlds must be properly aligned with respect to each other, or the illusion that the two worlds coexist will be compromised.

As a large number of potential AR applications are interactive, real time pose computation is required. Although the registration problem has received a lot of attention in the computer vision community, the problem of real-time registration is still far from being a solved problem, especially for unstructured environments. Ideally, an AR system should work in all environments, without the need to prepare the scene ahead of time, independently of the variations in experimental conditions (lighting, weather condition,...) which may exist between the application and the time the model of the scene was acquired.

For several years, the MAGRIT project has been aiming at developing on-line and marker-less methods for camera pose computation. The main difficulty with on-line tracking is to ensure robustness of the process over time. For off-line processes, robustness is achieved by using spatial and temporal coherence of the considered sequence through move-matching techniques. To get robustness for open-loop systems, we have investigated various methods, ranging from statistical methods to the use of hybrid camera/sensor systems. Many of these methods are dedicated to piecewise-planar scenes and combine the advantage of move-matching methods and model-based methods. In order to reduce statistical fluctuations in viewpoint computation, which lead to unpleasant jittering or sliding effects, we have also developed model selection techniques which allow us to noticeably improve the visual impression and to reduce drift over time. Another line of research which has been considered in the team to improve the reliability and the robustness of pose algorithms is to combine the camera with another form of sensor in order to compensate for the shortcomings of each technology [1].

The success of pose computation over time largely depends on the quality of the matching stage at the initialization stage. Indeed, the current image may be very different from the appearances described in the model both on the geometrical and the photometric sides. Research is thus conducted in the team on the use of probabilistic methods to establish robust correspondences of features. The use of *a contrario* has been investigated to achieve this aim [6]. We especially addressed the complex case of matching in scenes with repeated patterns which are common in urban scenes. We are also investigating the problem of matching images taken from very different viewpoints which is central for the re-localization issue in AR. Within the context of a scene model acquired with structure from motion techniques, we are currently investigating the use of viewpoint simulation in order to allow successful pose computation even if the considered image is far from the positions used to build the model.

Recently, the issue of tracking deformable objects has gained importance in the team. This topic is mainly addressed in the context of medical applications through the design of bio-mechanical models guided by visual features [2]. We have successfully investigated the use of such models in laparoscopy, with a vascularized model of the liver and with an hyper-elastic model for tongue tracking in US images. However, these results have been obtained so far in relatively controlled environments, with non pathological cases. When clinical routine applications are to be considered, many parameters and considerations need to be taken into account. Among the problems that need to be solved are the model representation, the specification of the range of physical parameters and the need to enforce the robustness of the tracking with respect to outliers, which are common in the interventional context...

3.2. Image-based Modeling

Modeling the scene is a fundamental issue in AR for many reasons. First, pose computation algorithms often use a model of the scene or at least some 3D knowledge on the scene. Second, effective AR systems require a model of the scene to support interactions between the virtual and the real objects such as occlusions, lighting reflexions, contacts...in real time. Unlike pose computation which has to be computed in a sequential way, scene modeling can be considered as an off-line or an on-line problem depending on the requirements of the targeted application. Interactive in-situ modeling techniques have thus been developed with the aim to enable the user to define what is relevant at the time the model is being built during the application. On the other hand, we also proposed off-line multimodal techniques, mainly dedicated to AR medical applications, with the aim to obtain realistic and possibly dynamic models of organs suitable for real time simulation.

In-situ modeling
In situ modeling allows a user to directly build a 3D model of his/her surrounding environment and verify the geometry against the physical world in real time. This is of particular interest in using AR in unprepared environments or building scenes that have an ephemeral existences (e.g. a film set) or cannot be accessed frequently (e.g. a nuclear power plant). We have especially investigated two systems, one based on the image content only and the other based on multiple data coming from different sensors (camera, inertial measurement unit, laser rangefinder). Both systems use the camera-mouse principle [5] (i.e. interactions are performed by aiming at the scene through a video camera) and both systems have been designed to acquire polygonal textured models, which are particularly useful for camera tracking and object insertion in AR.

Multimodal modeling for real time simulation
With respect to classical AR applications, AR in medical context differs in the nature and the size of the data which are available: a large amount of multimodal data is acquired on the patient or possibly on the operating room through sensing technologies or various image acquisitions. The challenge is to analyze these data, to extract interesting features, to fuse and to visualize this information in a proper way. Within the MAGRIT team, we address several key problems related to medical augmented environments. Being able to acquire multimodal data which are temporally synchronized and spatially registered is the first difficulty we face when considering medical AR. Another key requirement of AR medical systems is the availability of 3D (+t) models of the organ/patient built from images, to be overlaid onto the users's view of the environment.

Methods for multimodal modeling are strongly dependent on the image modalities and the organ specificities. We thus only address a restricted number of medical applications –interventional neuro-radiology, laparoscopic surgery, Augmented Head project– for which we have a strong expertise and close relationships with motivated clinicians. In these applications, our aim is to produce realistic models and then realistic simulations of the patient to be used for surgeon's training or patient's re-education/learning.

One of our main applications is about neuroradiology. For the last 20 years, we have been working in close collaboration with the neuroradiology laboratory (CHU-University Hospital of Nancy) and GE Healthcare. As several imaging modalities are now available in a intraoperative context (2D and 3D angiography, MRI, ...), our aim is to develop a multi-modality framework to help therapeutic decision and treatment.

We have mainly been interested in the effective use of a multimodality framework in the treatment of arteriovenous malformations (AVM) and aneurysms in the context of interventional neuroradiology. The goal of interventional gestures is to guide endoscopic tools towards the pathology with the aim to perform embolization of the AVM or to fill the aneurysmal cavity by placing coils. We have proposed and developed multimodality and augmented reality tools which make various image modalities (2D and 3D angiography, fluoroscopic images, MRI, ...) cooperate in order to help physicians in clinical routine. One of the successes of this collaboration is the implementation of the concept of *augmented fluoroscopy*, which helps the surgeon to guide endoscopic tools towards the pathology. Lately, in cooperation with the EPC SHACRA, we have proposed new methods for implicit modeling of the aneurysms with the aim of obtaining near real time simulation of the coil deployment in the aneurysm [8]. These works open the way towards near real time patient-based simulations of interventional gestures both for training or for planning.

3.3. Parameter estimation

Many problems in computer vision or image analysis can be formulated in terms of parameter estimation from image-based measurements. This is the case of many problems addressed in the team such as pose computation or image-guided estimation of 3D deformable models... Often traditional robust techniques which take into account the covariance on the measurements are sufficient to achieve reliable parameter estimation. However, depending on their number, their spatial distribution and the uncertainty on these measurements, some problems are very sensitive to noise and there is a considerable interest in considering how parameter estimation could be improved if additional information on the noise is available. Another common problem in our field of research is the need to estimate constitutive parameters of the models, such as (bio)-mechanical parameters for instance. Direct measurement methods are destructive and elaborating image based methods is thus highly desirable. Besides designing appropriate estimation algorithms, a fundamental question is to understand what group of parameters under study can be reliably estimated from a given experimental setup.

This line of research is relatively new in the team. One of the challenges is to improve image-based parameter estimation techniques considering sensor noise and specific image formation models. In a collaboration with the Pascal Institute (Clermont Ferrand), metrological performance enhancement for experimental solid mechanics has been addressed through the development of dedicated signal processing methods [12]. In the medical field, specific methods based on an adaptive evolutionary optimization strategy have been designed for estimating respiratory parameters [7]. In the context of designing realistic simulators for neuroradiology, we are now considering how parameters involved in the simulation could be adapted to fit real images.

MAIA Project-Team

3. Research Program

3.1. Sequential Decision Making

3.1.1. Synopsis and Research Activities

Sequential decision making consists, in a nutshell, in controlling the actions of an agent facing a problem whose solution requires not one but a whole sequence of decisions. This kind of problem occurs in a multitude of forms. For example, important applications addressed in our work include: Robotics, where the agent is a physical entity moving in the real world; Medicine, where the agent can be an analytic device recommending tests and/or treatments; Computer Security, where the agent can be a virtual attacker trying to identify security holes in a given network; and Business Process Management, where the agent can provide an auto-completion facility helping to decide which steps to include into a new or revised process. Our work on such problems is characterized by three main lines of research:

- (A) *Understanding how, and to what extent, to best model the problems.*
- (B) *Developing algorithms solving the problems and understanding their behavior.*
- (C) *Applying our results to complex applications.*

Before we describe some details of our work, it is instructive to understand the basic forms of problems we are addressing. We characterize problems along the following main dimensions:

- (1) Extent of the model: full vs. partial vs. none. This dimension concerns how complete we require the model of the problem – if any – to be. If the model is incomplete, then learning techniques are needed along with the decision making process.
- (2) Form of the model: factored vs. enumerative. Enumerative models explicitly list all possible world states and the associated actions etc. Factored models can be exponentially more compact, describing states and actions in terms of their behavior with respect to a set of higher-level variables.
- (3) World dynamics: deterministic vs. stochastic. This concerns our initial knowledge of the world the agent is acting in, as well as the dynamics of actions: is the outcome known a priori or are several outcomes possible?
- (4) Observability: full vs. partial. This concerns our ability to observe what our actions actually do to the world, i.e., to observe properties of the new world state. Obviously, this is an issue only if the world dynamics are stochastic.

These dimensions are wide-spread in the AI literature and are not exhaustive, in particular the MAIA team is also interested by discrete/continuous or centralized/decentralized problems. The complexity of solving a problem – both in theory and in practice – depends heavily on where it resides in this categorization. A common practice is to address simplified problems, leading to perhaps *sub-optimal* solutions while trying to characterize how far from the *optimal* solution we stand.

In what follows, we outline the main formal frameworks on which our work is based; while doing so, we highlight in a little more detail our core research questions. We then give a brief summary of how our work fits into the global research context.

3.1.2. Formal Frameworks

3.1.2.1. Deterministic Sequential Decision Making

Sequential decision making with deterministic world dynamics is most commonly known as *planning*, or *classical planning* [49]. Obviously, in such a setting every world state needs to be considered at most once, and thus enumerative models do not make sense (the problem description would have the same size as the space of possibilities to be explored). Planning approaches support factored description languages in which complex problems can be modeled in a compact way. Approaches to automatically learn such factored models do exist, however most works – and also most of our works on this form of sequential decision making – assume that the model is provided by the user of the planning technology. Formally, a problem instance, commonly referred to as a *planning task*, is a four-tuple $\langle V, A, I, G\rangle$. Here, V is a set of variables; a value assignment to the variables is a world state. A is a set of actions described in terms of two formulas over V: their preconditions and effects. I is the initial state, and G is a goal condition (again a formula over V). A solution, commonly referred to as a *plan*, is a schedule of actions that is applicable to I and achieves G.

Planning is *PSPACE-complete* even under strong restrictions on the formulas allowed in the planning task description. Research thus revolves around the development and understanding of search methods, which explore, in a variety of different ways, the space of possible action schedules. A particularly successful approach is *heuristic search*, where search is guided by information obtained in an automatically designed *relaxation* (simplified version) of the task. We investigate the design of relaxations, the connections between such design and the search space topology, and the construction of effective *planning systems* that exhibit good practical performance across a wide range of different inputs. Other important research lines concern the application of ideas successful in planning to stochastic sequential decision making (see next), and the development of technology supporting the user in model design.

3.1.2.2. Stochastic Sequential Decision Making

Markov Decision Processes (*MDP*) [51] are a natural framework for stochastic sequential decision making. An MDP is a four-tuple $\langle S, A, T, r\rangle$, where S is a set of states, A is a set of actions, $T(s, a, s') = P(s'|s, a)$ is the probability of transitioning to s' given that action a was chosen in state s, and $r(s, a, s')$ is the (possibly stochastic) reward obtained from taking action a in state s, and transitioning to state s'. In this framework, one looks for a *strategy*: a precise way for specifying the sequence of actions that induces, on average, an optimal sum of discounted rewards $\mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$. Here, $(r_0, r_1, ...)$ is the infinitely-long (random) sequence of rewards induced by the strategy, and $\gamma \in (0, 1)$ is a discount factor putting more weight on rewards obtained earlier. Central to the MDP framework is the Bellman equation, which characterizes the *optimal value function*V^*:

$$\forall s \in S, \quad V^*(s) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')[r(s, a, s') + \gamma V^*(s')].$$

Once the optimal value function is computed, it is straightforward to derive an optimal strategy, which is deterministic and memoryless, i.e., a simple mapping from states to actions. Such a strategy is usually called a *policy*. An *optimal policy* is any policy π^* that is *greedy* with respect to V^*, i.e., which satisfies:

$$\forall s \in S, \quad \pi(s) \in \arg\max_{a \in A} \sum_{s' \in S} T(s, a, s')[r(s, a, s') + \gamma V^*(s')].$$

An important extension of MDPs, known as Partially Observable MDPs (*POMDPs*) allows to account for the fact that the state may not be fully available to the decision maker. While the goal is the same as in an MDP (optimizing the expected sum of discounted rewards), the solution is more intricate. Any POMDP can be seen to be equivalent to an MDP defined on the space of probability distributions on states, called *belief states*. The Bellman-machinery then applies to the belief states. The specific structure of the resulting MDP makes it possible to iteratively approximate the optimal value function – which is convex in the *belief space* – by piecewise linear functions, and to deduce an optimal policy that maps belief states to actions. A further

extension, known as a DEC-POMDP, considers $n \geq 2$ agents that need to control the state dynamics in a decentralized way without direct communication.

The MDP model described above is enumerative, and the complexity of computing the optimal value function is *polynomial* in the size of that input. However, in examples of practical size, that complexity is still too high so naïve approaches do not scale. We consider the following situations: (i) when the state space is large, we study approximation techniques from both a theoretical and practical point of view; (ii) when the model is unknown, we study how to learn an optimal policy from samples (this problem is also known as Reinforcement Learning [55]); (iii) in factored models, where MDP models are a strict generalization of classical planning – and are thus at least *PSPACE*-hard to solve – we consider using search heuristics adapted from such (classical) planning.

Solving a POMDP is *PSPACE*-hard even given an enumerative model. In this framework, we are mainly looking for assumptions that could be exploited to reduce the complexity of the problem at hand, for instance when some actions have no effect on the state dynamics (*active sensing*). The decentralized version, DEC-POMDP, induces a significant increase in complexity (*NEXP*-complete). We tackle the challenging – even for (very) small state spaces – exact computation of finite-horizon optimal solutions through alternative reformulations of the problem. We also aim at proposing advanced heuristics to efficiently address problems with more agents and a longer time horizon.

3.2. Understanding and mastering complex systems

3.2.1. General context

There exist numerous examples of natural and artificial systems where self-organization and emergence occur. Such systems are composed of a set of simple entities interacting in a shared environment and exhibit complex collective behaviors resulting from the interactions of the local (or individual) behaviors of these entities. The properties that they exhibit, for instance robustness, explain why their study has been growing, both in the academic and the industrial field. They are found in a wide panel of fields such as sociology (opinion dynamics in social networks), ecology (population dynamics), economy (financial markets, consumer behaviors), ethology (swarm intelligence, collective motion), cellular biology (cells/organ), computer networks (ad-hoc or P2P networks), etc.

More precisely, the systems we are interested in are characterized by:

- *locality*: Elementary components have only a partial perception of the system's state, similarly, a component can only modify its surrounding environment.
- *individual simplicity*: components have a simple behavior, in most cases it can be modeled by stimulus/response laws or by look-up tables. One way to estimate this simplicity is to count the number of stimulus/response rules for instance.
- *emergence*: It is generally difficult to predict the global behavior of the system from the local individual behaviors. This difficulty of prediction is often observed empirically and in some cases (e.g., cellular automata) one can show that the prediction of the global properties of a system is an undecidable problem. However, observations coming from simulations of the system may help us to find the regularities that occur in the system's behavior (even in a probabilistic meaning). Our interest is to work on problems where a full mathematical analysis seems out of reach and where it is useful to observe the system with large simulations. In return, it is frequent that the properties observed empirically are then studied on an analytical basis. This approach should allow us to understand where lies the frontier between simulation and analysis.
- *levels of description and observation*: Describing a complex system involves at least two levels: the micro level that regards how a component behaves, and the macro level associated with the collective behavior. Usually, understanding a complex system requires to link the description of a component behavior with the observation of a collective phenomenon: establishing this link may require various levels, which can be obtained only with a careful analysis of the system.

We now describe the type of models that are studied in our group.

3.2.2. Multi-agent models

We represent these complex systems with reactive multi-agent systems (RMAS). Multi-agent systems are defined by a set of reactive agents, an environment, a set of interactions between agents and a resulting organization. They are characterized by a decentralized control shared among agents: each agent has an internal state, has access to local observations and influences the system through stimulus response rules. Thus, the collective behavior results from individual simplicity and successive actions and interactions of agents through the environment.

Reactive multi-agent systems present several advantages for modeling complex systems

- agents are explicitly represented in the system and have the properties of local action, interaction and observation;
- each agent can be described regardless of the description of the other agents, multi-agent systems allow explicit heterogeneity among agents which is often at the root of collective emergent phenomena;
- multi-agent systems can be executed through simulation and provide good models to investigate the complex link between global and local phenomena for which analytic studies are hard to perform.

By proposing two different levels of description, the local level of the agents and the global level of the phenomenon, and several execution models, multi-agent systems constitute an interesting tool to study the link between local and global properties.

Despite a widespread use of multi-agent systems, their framework still needs many improvements to be fully accessible to computer scientists from various backgrounds. For instance, there is no generic model to mathematically define a reactive multi-agent system and to describe its interactions. This situation is in contrast with the field of cellular automata, for instance, and underlines that a unification of multi-agent systems under a general framework is a question that still remains to be tackled. We now list the different challenges that, in part, contribute to such an objective.

3.2.3. Current challenges

Our work is structured around the following challenges that combine both theoretical and experimental approaches.

3.2.3.1. Providing formal frameworks

A widespread and consensual formal definition of a multi-agent system is lacking. Our research aims at translating the concepts from the field of complex systems into the multi-agent systems framework.

One objective of this research is to remove the potential ambiguities that can appear if one describes a system without explicitly formulating each aspect of the simulation framework. As a benefit, the reproduction of experiments is facilitated. Moreover, this approach is intended to gain a better insight of the self-organization properties of the systems.

Another important question consists in monitoring the evolution of complex systems. Our objective is to provide some quantitative characteristics of the system such as local or global stability, robustness, complexity, etc. Describing our models as dynamical systems leads us to use specific tools of this mathematical theory as well as statistical tools.

3.2.3.2. Controlling complex dynamical system

Since there is no central control of our systems, one question of interest is to know under which conditions it is possible to guarantee a given property when the system is subject to perturbations. We tackle this issue by designing exogenous control architectures where control actions are envisaged as perturbations in the system. As a consequence, we seek to develop control mechanisms that can change the global behavior of a system without modifying the agent behavior (and not violating the autonomy property).

3.2.3.3. Designing systems

The aim is to design individual behaviors and interactions in order to produce a desired collective output. This output can be a collective pattern to reproduce in case of simulation of natural systems. In that case, from individual behaviors and interactions we study if (and how) the collective pattern is produced. We also tackle "inverse problems" (decentralized gathering problem, density classification problem, etc.) which consist in finding individual behaviors in order to solve a given problem.

MANAO Project-Team

3. Research Program

3.1. Related Scientific Domains

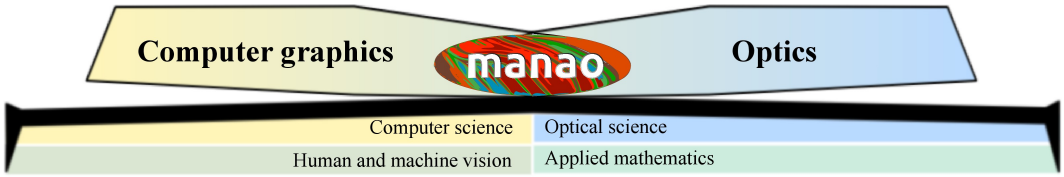


Figure 3. Related scientific domains of the MANAO project.

The *MANAO* project aims to study, acquire, model, and render the interactions between the three components that are light, shape, and matter from the viewpoint of an observer. As detailed more lengthily in the next section, such a work will be done using the following approach: first, we will tend to consider that these three components do not have strict frontiers when considering their impacts on the final observers; then, we will not only work in **computer graphics**, but also at the intersections of computer graphics and **optics**, exploring the mutual benefits that the two domains may provide. It is thus intrinsically a **transdisciplinary** project (as illustrated in Figure 3) and we expect results in both domains.

Thus, the proposed team-project aims at establishing a close collaboration between computer graphics (e.g., 3D modeling, geometry processing, shading techniques, vector graphics, and GPU programming) and optics (e.g., design of optical instruments, and theories of light propagation). The following examples illustrate the strengths of such a partnership. First, in addition to simpler radiative transfer equations [50] commonly used in computer graphics, research in the later will be based on state-of-the-art understanding of light propagation and scattering in real environments. Furthermore, research will rely on appropriate instrumentation expertise for the measurement [62], [63] and display [61] of the different phenomena. Reciprocally, optics researches may benefit from the expertise of computer graphics scientists on efficient processing to investigate interactive simulation, visualization, and design. Furthermore, new systems may be developed by unifying optical and digital processing capabilities. Currently, the scientific background of most of the team members is related to computer graphics and computer vision. A large part of their work have been focused on simulating and analyzing optical phenomena as well as in acquiring and visualizing them. Combined with the close collaboration with the optics laboratory (LP2N) and with the students issued from the “Institut d’Optique”, this background ensures that we can expect the following results from the project: the construction of a common vocabulary for tightening the collaboration between the two scientific domains and creating new research topics. By creating this context, we expect to attract (and even train) more trans-disciplinary researchers.

At the boundaries of the *MANAO* project lie issues in **human and machine vision**. We have to deal with the former whenever a human observer is taken into account. On one side, computational models of human vision are likely to guide the design of our algorithms. On the other side, the study of interactions between light, shape, and matter may shed some light on the understanding of visual perception. The same kind of connections are expected with machine vision. On the one hand, traditional computational methods for acquisition (such as photogrammetry) are going to be part of our toolbox. On the other hand, new display technologies (such as augmented reality) are likely to benefit from our integrated approach and systems. In the *MANAO* project we are mostly users of results from human vision. When required, some experimentation

might be done in collaboration with experts from this domain, like with the European PRISM project (cf. Section TODO). For machine vision, provided the tight collaboration between optical and digital systems, research will be carried out inside the *MANAO* project.

Analysis and modeling rely on **tools from applied mathematics** such as differential and projective geometry, multi-scale models, frequency analysis [52] or differential analysis [86], linear and non-linear approximation techniques, stochastic and deterministic integrations, and linear algebra. We not only rely on classical tools, but also investigate and adapt recent techniques (e.g., improvements in approximation techniques), focusing on their ability to run on modern hardware: the development of our own tools (such as Eigen, see Section 4.1.2) is essential to control their performances and their abilities to be integrated into real-time solutions or into new instruments.

3.2. Research axes

The *MANAO* project is organized around four research axes that cover the large range of expertise of its members and associated members. We briefly introduce these four axes in this section. More details and their inter-influences that are illustrated in the Figure 2 will be given in the following sections.

Axis 1 is the theoretical foundation of the project. Its main goal is to increase the understanding of light, shape, and matter interactions by combining expertise from different domains: optics and human/machine vision for the analysis and computer graphics for the simulation aspect. The goal of our analyses is to identify the different layers/phenomena that compose the observed signal. In a second step, the development of physical simulations and numerical models of these identified phenomena is a way to validate the pertinence of the proposed decompositions.

In Axis 2, the final observers are mainly physical captors. Our goal is thus the development of new acquisition and display technologies that combine optical and digital processes in order to reach fast transfers between real and digital worlds, in order to increase the convergence of these two worlds.

Axes 3 and 4 focus on two aspects of computer graphics: rendering, visualization and illustration in Axis 3, and editing and modeling (content creation) in Axis 4. In these two axes, the final observers are mainly human users, either generic users or expert ones (e.g., archaeologist [91], computer graphics artists).

3.3. Axis 1: Analysis and Simulation

Challenge: Definition and understanding of phenomena resulting from interactions between light, shape, and matter as seen from an observer point of view.

Results: Theoretical tools and numerical models for analyzing and simulating the observed optical phenomena.

To reach the goals of the *MANAO* project, we need to **increase our understanding** of how light, shape, and matter act together in synergy and how the resulting signal is finally observed. For this purpose, we need to identify the different phenomena that may be captured by the targeted observers. This is the main objective of this research axis, and it is achieved by using three approaches: the simulation of interactions between light, shape, and matter, their analysis and the development of new numerical models. This resulting improved knowledge is a foundation for the researches done in the three other axes, and the simulation tools together with the numerical models serve the development of the joint optical/digital systems in Axis 2 and their validation.

One of the main and earliest goals in computer graphics is to faithfully reproduce the real world, focusing mainly on light transport. Compared to researchers in physics, researchers in computer graphics rely on a subset of physical laws (mostly radiative transfer and geometric optics), and their main concern is to efficiently use the limited available computational resources while developing as fast as possible algorithms. For this purpose, a large set of tools has been introduced to take a **maximum benefit of hardware** specificities. These tools are often dedicated to specific phenomena (e.g., direct or indirect lighting, color bleeding, shadows, caustics). An efficiency-driven approach needs such a classification of light paths [58] in order to develop tailored strategies [104]. For instance, starting from simple direct lighting, more complex phenomena have

been progressively introduced: first diffuse indirect illumination [56], [95], then more generic inter-reflections [65], [50] and volumetric scattering [92], [47]. Thanks to this search for efficiency and this classification, researchers in computer graphics have developed a now recognized expertise in fast-simulation of light propagation. Based on finite elements (radiosity techniques) or on unbiased Monte Carlo integration schemes (ray-tracing, particle-tracing, ...), the resulting algorithms and their combination are now sufficiently accurate to be used-back in physical simulations. The *MANAO* project will continue the search for **efficient and accurate simulation** techniques, but extending it from computer graphics to optics. Thanks to the close collaboration with scientific researchers from optics, new phenomena beyond radiative transfer and geometric optics will be explored.

Search for algorithmic efficiency and accuracy has to be done in parallel with **numerical models**. The goal of visual fidelity (generalized to accuracy from an observer point of view in the project) combined with the goal of efficiency leads to the development of alternative representations. For instance, common classical finite-element techniques compute only basis coefficients for each discretization element: the required discretization density would be too large and to computationally expensive to obtain detailed spatial variations and thus visual fidelity. Examples includes texture for decorrelating surface details from surface geometry and high-order wavelets for a multi-scale representation of lighting [46]. The numerical complexity explodes when considering directional properties of light transport such as radiance intensity (Watt per square meter and per steradian - $W.m^{-2}.sr^{-1}$), reducing the possibility to simulate or accurately represent some optical phenomena. For instance, Haar wavelets have been extended to the spherical domain [94] but are difficult to extend to non-piecewise-constant data [97]. More recently, researches prefer the use of Spherical Radial Basis Functions [100] or Spherical Harmonics [85]. For more complex data, such as reflective properties (e.g., BRDF [79], [66] - 4D), ray-space (e.g., Light-Field [76] - 4D), spatially varying reflective properties (6D - [89]), new models, and representations are still investigated such as rational functions [82] or dedicated models [33] and parameterizations [93], [98]. For each (newly) defined phenomena, we thus explore the space of possible numerical representations to determine the **most suited one for a given application**, like we have done for BRDF [82].

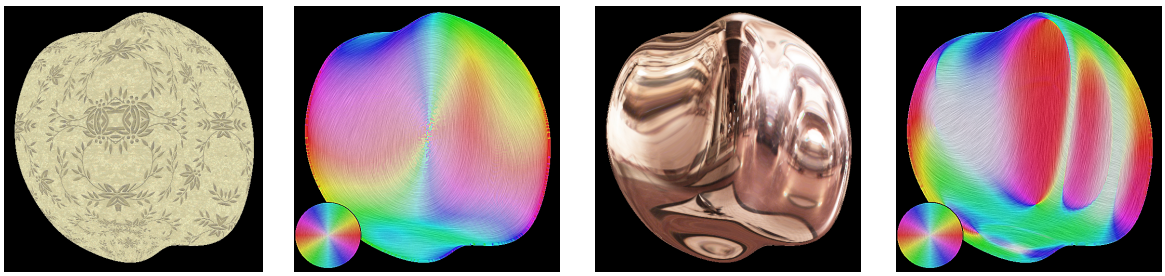


Figure 4. First-oder analysis [105] have shown that shading variations are caused by depth variations (first-order gradient field) and by normal variations (second-order fields). These fields are visualized using hue and saturation to indicate direction and magnitude of the flow respectively.

Before being able to simulate or to represent the different **observed phenomena**, we need to define and describe them. To understand the difference between an observed phenomenon and the classical light, shape, and matter decomposition, we can take the example of a highlight. Its observed shape (by a human user or a sensor) is the resulting process of the interaction of these three components, and can be simulated this way. However, this does not provide any intuitive understanding of their relative influence on the final shape: an artist will directly describe the resulting shape, and not each of the three properties. We thus want to decompose the observed signal into models for each scale that can be easily understandable, representable,

and manipulable. For this purpose, we will rely on the **analysis** of the resulting interaction of light, shape, and matter as observed by a human or a physical sensor. We first consider this analysis from an **optical point of view**, trying to identify the different phenomena and their scale according to their mathematical properties (e.g., differential [86] and frequency analysis [52]). Such an approach has leaded us to exhibit the influence of surfaces flows (depth and normal gradients) into lighting pattern deformation (see Figure 4). For a **human observer**, this correspond to one recent trend in computer graphics that takes into account the human visual systems [53] both to evaluate the results and to guide the simulations.

3.4. Axis 2: From Acquisition to Display

Challenge: Convergence of optical and digital systems to blend real and virtual worlds.

Results: Instruments to acquire real world, to display virtual world, and to make both of them interact.

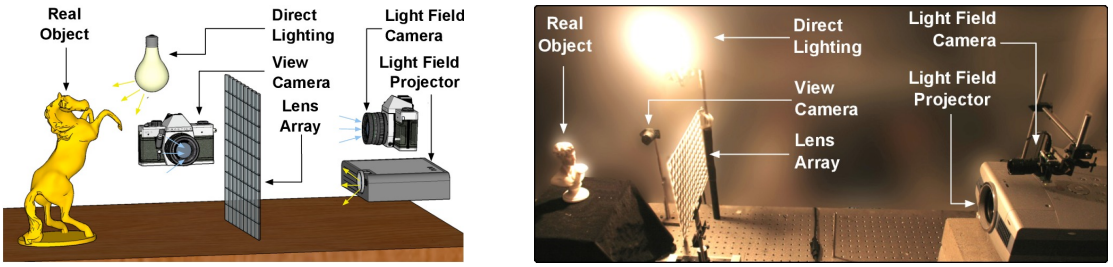


Figure 5. Light-Field transfer: global illumination between real and synthetic objects [45]

For this axis, we investigate *unified acquisition and display systems*, that is systems which combine optical instruments with digital processing. From digital to real, we investigate new display approaches [76], [61]. We consider projecting systems and surfaces [42], for personal use, virtual reality and augmented reality [36]. From the real world to the digital world, we favor direct measurements of parameters for models and representations, using (new) optical systems unless digitization is required [55], [54]. These resulting systems have to acquire the different phenomena described in Axis 1 and to display them, in an efficient manner [59], [34], [60], [63]. By efficient, we mean that we want to shorten the path between the real world and the virtual world by increasing the data bandwidth between the real (analog) and the virtual (digital) worlds, and by reducing the latency for real-time interactions (we have to prevent unnecessary conversions, and to reduce processing time). To reach this goal, the systems have to be designed as a whole, not by a simple concatenation of optical systems and digital processes, nor by considering each component independently [64].

To increase data bandwidth, one solution is to **parallelize more and more the physical systems**. One possible solution is to multiply the number of simultaneous acquisitions (e.g., simultaneous images from multiple viewpoints [63], [84]). Similarly, increasing the number of viewpoints is a way toward the creation of full 3D displays [76]. However, full acquisition or display of 3D real environments theoretically requires a continuous field of viewpoints, leading to huge data size. Despite the current belief that the increase of computational power will fill the missing gap, when it comes to visual or physical realism, if you double the processing power, people may want four times more accuracy, thus increasing data size as well. Furthermore, this leads to solutions that are not energy efficient and thus cannot be embedded into mobile devices. To reach the best performances, a trade-off has to be found between the amount of data required to represent accurately the reality and the amount of required processing. This trade-off may be achieved using **compressive sensing**. Compressive sensing is a new trend issued from the applied mathematics community that provides tools to accurately reconstruct a signal from a small set of measurements assuming that it is sparse in a transform domain (e.g., [83], [108]).

We prefer to achieve this goal by avoiding as much as possible the classical approach where acquisition is followed by a fitting step: this requires in general a large amount of measurements and the fitting itself may consume consequently too much memory and preprocessing time. By **preventing unnecessary conversion** through fitting techniques, such an approach increase the speed and reduce the data transfer for acquisition but also for display. One of the best recent examples is the work of Cossairt et al. [45]. The whole system is designed around a unique representation of the energy-field issued from (or leaving) a 3D object, either virtual or real: the Light-Field. A Light-Field encodes the light emitted in any direction from any position on an object. It is acquired thanks to a lens-array that leads to the capture of, and projection from, multiple simultaneous viewpoints. A unique representation is used for all the steps of this system. Lens-arrays, parallax barriers, and coded-aperture [71] are one of the key technologies to develop such acquisition (e.g., Light-Field camera [0] [64] and acquisition of light-sources [55]), projection systems (e.g., auto-stereoscopic displays). Such an approach is versatile and may be applied to improve classical optical instruments [69]. More generally, by designing unified optical and digital systems [80], it is possible to leverage the requirement of processing power, the memory footprint, and the cost of optical instruments.

Those are only some examples of what we investigate. We also consider the following approaches to develop new unified systems. First, similar to (and based on) the analysis goal of Axis 1, we have to take into account as much as possible the characteristics of the measurement setup. For instance, when fitting cannot be avoided, integrating them may improve both the processing efficiency and accuracy [82]. Second, we have to integrate signals from multiple sensors (such as GPS, accelerometer, ...) to prevent some computation (e.g., [72]). Finally, the experience of the group in surface modeling help the design of optical surfaces [67] for light sources or head-mounted displays.

3.5. Axis 3: Rendering, Visualization and Illustration

Challenge: How to offer the most legible signal to the final observer in real-time?

Results: High-level shading primitives, expressive rendering techniques for object depiction, real-time realistic rendering algorithms

(a) Global illumination [81] (b) Shadows [35] (c) Shape enhancement [10] (d) Shape depiction [32]

Figure 6. In the MANAO project, we are investigating rendering techniques from realistic solutions (e.g., inter-reflections (a) and shadows (b)) to more expressive ones (shape enhancement (c) with realistic style and shape depiction (d) with stylized style) for visualization.

The main goal of this axis is to offer to the final observer, in this case mostly a human user, the most legible signal in real-time. Thanks to the analysis and to the decomposition in different phenomena resulting from interactions between light, shape, and matter (Axis 1), and their perception, we can use them to convey essential information in the most pertinent way. Here, the word *pertinent* can take various forms depending on the application.

[0] Lytro, http://www.lytro.com/

In the context of scientific illustration and visualization, we are primarily interested in tools to convey shape or material characteristics of objects in animated 3D scenes. **Expressive rendering** techniques (see Figure 6 c,d) provide means for users to depict such features with their own style. To introduce our approach, we detail it from a shape-depiction point of view, domain where we have acquired a recognized expertise. Prior work in this area mostly focused on stylization primitives to achieve line-based rendering [106], [68] or stylized shading [40],[10] with various levels of abstraction. A clear representation of important 3D **object features** remains a major challenge for better shape depiction, stylization and abstraction purposes. Most existing representations provide only local properties (e.g., curvature), and thus lack characterization of broader shape features. To overcome this limitation, we are developing higher level descriptions of shape [31] with increased robustness to sparsity, noise, and outliers. This is achieved in close collaboration with Axis 1 by the use of higher-order local fitting methods, multi-scale analysis, and global regularization techniques. In order not to neglect the observer and the material characteristics of the objects, we couple this approach with an analysis of the appearance model. To our knowledge, this is an approach which has not been considered yet. This research direction is at the heart of the *MANAO* project, and has a strong connection with the analysis we plan to conduct in Axis 1. Material characteristics are always considered at the light ray level, but an understanding of **higher-level primitives** (like the shape of highlights and their motion) would help us to produce more legible renderings and permit novel stylizations; for instance, there is no method that is today able to create stylized renderings that follow the motion of highlights or shadows. We also believe such tools also play a fundamental role for geometry processing purposes (such as shape matching, reassembly, simplification), as well as for editing purposes as discussed in Axis 4.

In the context of **real-time photo-realistic rendering** ((see Figure 6 a,b), the challenge is to compute the most plausible images with minimal effort. During the last decade, a lot of work has been devoted to design approximate but real-time rendering algorithms of complex lighting phenomena such as soft-shadows [107], motion blur [52], depth of field [96], reflexions, refractions, and inter-reflexions. For most of these effects it becomes harder to discover fundamentally new and faster methods. On the other hand, we believe that significant speedup can still be achieved through more clever use of **massively parallel architectures** of the current and upcoming hardware, and/or through more clever tuning of the current algorithms. In particular, regarding the second aspect, we remark that most of the proposed algorithms depend on several parameters which can be used to **trade the speed over the quality**. Significant speed-up could thus be achieved by identifying effects that would be masked or facilitated and thus devote appropriate computational resources to the rendering [70], [51]. Indeed, the algorithm parameters controlling the quality vs speed are numerous without a direct mapping between their values and their effect. Moreover, their ideal values vary over space and time, and to be effective such an auto-tuning mechanism has to be extremely fast such that its cost is largely compensated by its gain. We believe that our various work on the analysis of the appearance such as in Axis 1 could be beneficial for such purpose too.

Realistic and real-time rendering is closely related to Axis 2: real-time rendering is a requirement to close the loop between real world and digital world. We have to thus develop algorithms and rendering primitives that allow the integration of the acquired data into real-time techniques. We have also to take care of that these real-time techniques have to work with new display systems. For instance, stereo, and more generally multi-view displays are based on the multiplication of simultaneous images. Brute force solutions consist in independent rendering pipeline for each viewpoint. A more energy-efficient solution would take advantages of the computation parts that may be factorized. Another example is the rendering techniques based on image processing, such as our work on augmented reality [44]. Independent image processing for each viewpoint may disturb the feeling of depth by introducing inconsistent information in each images. Finally, more dedicated displays [61] would require new rendering pipelines.

3.6. Axis 4: Editing and Modeling

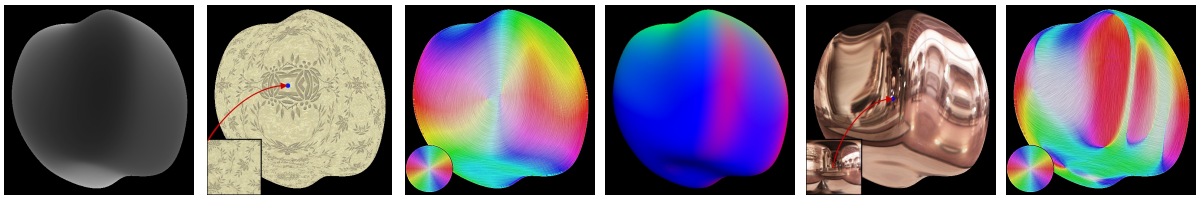
Challenge: Editing and modeling appearance using drawing- or sculpting-like tools through high level representations.

Results: High-level primitives and hybrid representations for appearance and shape.

During the last decade, the domain of computer graphics has exhibited tremendous improvements in image quality, both for 2D applications and 3D engines. This is mainly due to the availability of an ever increasing amount of shape details, and sophisticated appearance effects including complex lighting environments. Unfortunately, with such a growth in visual richness, even so-called *vectorial* representations (e.g., subdivision surfaces, Bézier curves, gradient meshes, etc.) become very dense and unmanageable for the end user who has to deal with a huge mass of control points, color labels, and other parameters. This is becoming a major challenge, with a necessity for novel representations. This Axis is thus complementary of Axis 3: the focus is the development of primitives that are easy to use for modeling and editing.

More specifically, we plan to investigate *vectorial representations* that would be amenable to the production of rich shapes with a minimal set of primitives and/or parameters. To this end we plan to build upon our insights on dynamic local reconstruction techniques and implicit surfaces [3] [39]. When working in 3D, an interesting approach to produce detailed shapes is by means of procedural geometry generation. For instance, many natural phenomena like waves or clouds may be modeled using a combination of procedural functions. Turning such functions into triangle meshes (main rendering primitives of GPUs) is a tedious process that appears not to be necessary with an adapted vectorial shape representation where one could directly turn procedural functions into implicit geometric primitives. Since we want to prevent unnecessary conversions in the whole pipeline (here, between modeling and rendering steps), we will also consider *hybrid representations* mixing meshes and implicit representations. Such research has thus to be conducted while considering the associated editing tools as well as performance issues. It is indeed important to keep *real-time performance* (cf. Axis 2) throughout the interaction loop, from user inputs to display, via editing and rendering operations. Finally, it would be interesting to add *semantic information* into 2D or 3D geometric representations. Semantic geometry appears to be particularly useful for many applications such as the design of more efficient manipulation and animation tools, for automatic simplification and abstraction, or even for automatic indexing and searching. This constitutes a complementary but longer term research direction.

In the *MANAO* project, we want to investigate representations beyond the classical light, shape, and matter decomposition. We thus want to directly control the appearance of objects both in 2D and 3D applications (e.g., [102]): this is a core topic of computer graphics. When working with 2D vector graphics, digital artists must carefully set up color gradients and textures: examples range from the creation of 2D logos to the photo-realistic imitation of object materials. Classic vector primitives quickly become impractical for creating illusions of complex materials and illuminations, and as a result an increasing amount of time and skill is required. This is only for still images. For animations, vector graphics are only used to create legible appearances composed of simple lines and color gradients. There is thus a need for more complex primitives that are able to accommodate complex reflection or texture patterns, while keeping the ease of use of vector graphics. For instance, instead of drawing color gradients directly, it is more advantageous to draw flow lines that represent local surface concavities and convexities. Going through such an intermediate structure then allows to deform simple material gradients and textures in a coherent way (see Figure 7), and animate them all at once. The manipulation of 3D object materials also raises important issues. Most existing material models are tailored to faithfully reproduce physical behaviors, not to be *easily controllable* by artists. Therefore artists learn to tweak model parameters to satisfy the needs of a particular shading appearance, which can quickly become cumbersome as the complexity of a 3D scene increases. We believe that an alternative approach is required, whereby material appearance of an object in a typical lighting environment is directly input (e.g., painted or drawn), and adapted to match a plausible material behavior. This way, artists will be able to create their own appearance (e.g., by using our shading primitives [102]), and replicate it to novel illumination environments and 3D models. For this purpose, we will rely on the decompositions and tools issued from Axis 1.



(a) (b) (c) (d) (e) (f)

Figure 7. Based on our analysis [105] (Axis 1), we have designed a system that mimics texture (left) and shading (right) effects using image processing alone. It takes depth (a) and normal (d) images as input, and uses them to deform images (b-e) in ways that closely approximate surface flows (c-f). It provides a convincing, yet artistically controllable illusion of 3D shape conveyed through texture or shading cues.

MAVERICK Project-Team

3. Research Program

3.1. Introduction

The Maverick project-team aims at producing representations and algorithms for efficient, high-quality computer generation of pictures and animations through the study of four **research problems**:

- *Computer Visualization* where we take as input a large localized dataset and represent it in a way that will let an observer understand its key properties. Visualization can be used for data analysis, for the results of a simulation, for medical imaging data...
- *Expressive Rendering*, where we create an artistic representation of a virtual world. Expressive rendering corresponds to the generation of drawings or paintings of a virtual scene, but also to some areas of computational photography, where the picture is simplified in specific areas to focus the attention.
- *Illumination Simulation*, where we model the interaction of light with the objects in the scene, resulting in a photorealistic picture of the scene. Research include improving the quality and photorealism of pictures, including more complex effects such as depth-of-field or motion-blur. We are also working on accelerating the computations, both for real-time photorealistic rendering and offline, high-quality rendering.
- *Complex Scenes*, where we generate, manage, animate and render highly complex scenes, such as natural scenes with forests, rivers and oceans, but also large datasets for visualization. We are especially interested in interactive visualization of complex scenes, with all the associated challenges in terms of processing and memory bandwidth.

The fundamental research interest of Maverick is first, *understanding* what makes a picture useful, powerful and interesting for the user, and second *designing* algorithms to create and improve these pictures.

3.2. Research approaches

We will address these research problems through three interconnected research approaches:

3.2.1. Picture Impact

Our first research axis deals with the *impact* pictures have on the viewer, and how we can improve this impact. Our research here will target:

- *evaluating user response:* we need to evaluate how the viewers respond to the pictures and animations generated by our algorithms, through user studies, either asking the viewer about what he perceives in a picture or measuring how his body reacts (eye tracking, position tracking).
- *removing artefacts and discontinuities:* temporal and spatial discontinuities perturb viewer attention, distracting the viewer from the main message. These discontinuities occur during the picture creation process; finding and removing them is a difficult process.

3.2.2. Data Representation

The data we receive as input for picture generation is often unsuitable for interactive high-quality rendering: too many details, no spatial organisation... Similarly the pictures we produce or get as input for other algorithms can contain superfluous details.

One of our goals is to develop new data representations, adapted to our requirements for rendering. This includes fast access to the relevant information, but also access to the specific hierarchical level of information needed: we want to organize the data in hierarchical levels, pre-filter it so that sampling at a given level also gives information about the underlying levels. Our research for this axis include filtering, data abstraction, simplification and stylization.

The input data can be of any kind: geometric data, such as the model of an object, scientific data before visualization, pictures and photographs. It can be time-dependent or not; time-dependent data bring an additional level of challenge on the algorithm for fast updates.

3.2.3. Prediction and simulation

Our algorithms for generating pictures require computations: sampling, integration, simulation... These computations can be optimized if we already know the characteristics of the final picture. Our recent research has shown that it is possible to predict the local characteristics of a picture by studying the phenomena involved: the local complexity, the spatial variations, their direction...

Our goal is to develop new techniques for predicting the properties of a picture, and to adapt our image-generation algorithms to these properties, for example by sampling less in areas of low variation.

Our research problems and approaches are all cross-connected. Research on the *impact* of pictures is of interest in three different research problems: *Computer Visualization*, *Expressive rendering* and *Illumination Simulation*. Similarly, our research on *Illumination simulation* will use all three research approaches: impact, representations and prediction.

3.3. Cross-cutting research issues

Beyond the connections between our problems and research approaches, we are interested in several issues, which are present throughout all our research:

sampling is an ubiquitous process occurring in all our application domains, whether photorealistic rendering (*e.g.* photon mapping), expressive rendering (*e.g.* brush strokes), texturing, fluid simulation (Lagrangian methods), etc. When sampling and reconstructing a signal for picture generation, we have to ensure both coherence and homogeneity. By *coherence*, we mean not introducing spatial or temporal discontinuities in the reconstructed signal. By *homogeneity*, we mean that samples should be placed regularly in space and time. For a time-dependent signal, these requirements are conflicting with each other, opening new areas of research.

filtering is another ubiquitous process, occuring in all our application domains, whether in realistic rendering (*e.g.* for integrating height fields, normals, material properties), expressive rendering (*e.g.* for simplifying strokes), textures (through non-linearity and discontinuities). It is especially relevant when we are replacing a signal or data with a lower resolution (for hierarchical representation); this involves filtering the data with a reconstruction kernel, representing the transition between levels.

performance and scalability are also a common requirement for all our applications. We want our algorithms to be usable, which implies that they can be used on large and complex scenes, placing a great importance on scalability. For some applications, we target interactive and real-time applications, with an update frequency between 10 Hz and 120 Hz.

coherence and continuity in space and time is also a common requirement of realistic as well as expressive models which must be ensured despite contradictory requirements. We want to avoid flickering and aliasing.

animation: our input data is likely to be time-varying (*e.g.* animated geometry, physical simulation, time-dependent dataset). A common requirement for all our algorithms and data representation is that they must be compatible with animated data (fast updates for data structures, low latency algorithms...).

3.4. Methodology

Our research is guided by several methodological principles:

Experimentation: to find solutions and phenomenological models, we use experimentation, performing statistical measurements of how a system behaves. We then extract a model from the experimental data.

Validation: for each algorithm we develop, we look for experimental validation: measuring the behavior of the algorithm, how it scales, how it improves over the state-of-the-art... We also compare our algorithms to the exact solution. Validation is harder for some of our research domains, but it remains a key principle for us.

Reducing the complexity of the problem: the equations describing certain behaviors in image synthesis can have a large degree of complexity, precluding computations, especially in real time. This is true for physical simulation of fluids, tree growth, illumination simulation... We are looking for *emerging phenomena* and *phenomenological models* to describe them (see framed box "Emerging phenomena"). Using these, we simplify the theoretical models in a controlled way, to improve user interaction and accelerate the computations.

Transfering ideas from other domains: Computer Graphics is, by nature, at the interface of many research domains: physics for the behavior of light, applied mathematics for numerical simulation, biology, algorithmics... We import tools from all these domains, and keep looking for new tools and ideas.

Develop new fondamental tools: In situations where specific tools are required for a problem, we will proceed from a theoretical framework to develop them. These tools may in return have applications in other domains, and we are ready to disseminate them.

Collaborate with industrial partners: we have a long experiment of collaboration with industrial partners. These collaborations bring us new problems to solve, with short-term or medium-term transfert opportunities. When we cooperate with these partners, we have to find *what they need*, which can be very different from *what they want*, their expressed need.

MIMETIC Project-Team

3. Research Program

3.1. Biomechanics and Motion Control

Human motion control is a very complex phenomenon that involves several layered systems, as shown in Figure 3 . Each layer of this controller is responsible for dealing with perceptual stimuli in order to decide the actions that should be applied to the human body and his environment. Due to the intrinsic complexity of the information (internal representation of the body and mental state, external representation of the environment) used to perform this task, it is almost impossible to model all the possible states of the system. Even for simple problems, there generally exist infinity of solutions. For example, from the biomechanical point of view, there are much more actuators (i.e. muscles) than degrees of freedom leading to infinity of muscle activation patterns for a unique joint rotation. From the reactive point of view there exist infinity of paths to avoid a given obstacle in navigation tasks. At each layer, the key problem is to understand how people select one solution among these infinite state spaces. Several scientific domains have addressed this problem with specific points of view, such as physiology, biomechanics, neurosciences and psychology.

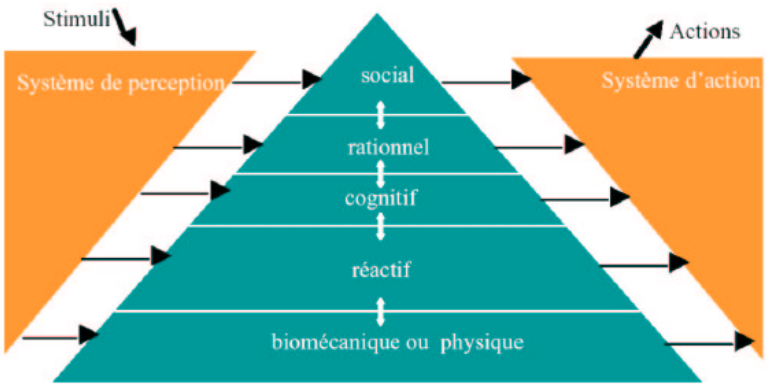


Figure 3. Layers of the motion control natural system in humans.

In biomechanics and physiology, researchers have proposed hypotheses based on accurate joint modeling (to identify the real anatomical rotational axes), energy minimization, force and torques minimization, comfort maximization (i.e. avoiding joint limits), and physiological limitations in muscle force production. All these constraints have been used in optimal controllers to simulate natural motions. The main problem is thus to define how these constraints are composed altogether such as searching the weights used to linearly combine these criteria in order to generate a natural motion. Musculoskeletal models are stereotyped examples for which there exist infinity of muscle activation patterns, especially when dealing with antagonist muscles. An unresolved problem is to define how using the above criteria to retrieve the actual activation patterns while optimization approaches still lead to unrealistic ones. It is still an open problem that will require multidisciplinary skills including computer simulation, constraint solving, biomechanics, optimal control, physiology and neurosciences.

In neuroscience, researchers have proposed other theories, such as coordination patterns between joints driven by simplifications of the variables used to control the motion. The key idea is to assume that instead of controlling all the degrees of freedom, people control higher level variables which correspond to combination of joint angles. In walking, data reduction techniques such as Principal Component Analysis have shown that lower-limb joint angles are generally projected on a unique plan whose angle in the state space is associated with energy expenditure. Although there exist knowledge on specific motion, such as locomotion or grasping, this type of approach is still difficult to generalize. The key problem is that many variables are coupled and it is very difficult to objectivly study the behavior of a unique variable in various motor tasks. Computer simulation is a promising method to evaluate such type of assumptions as it enables to accurately control all the variables and to check if it leads to natural movements.

Neurosciences also address the problem of coupling perception and action by providing control laws based on visual cues (or any other senses), such as determining how the optical flow is used to control direction in navigation tasks, while dealing with collision avoidance or interception. Coupling of the control variables is enhanced in this case as the state of the body is enriched by the big amount of external information that the subject can use. Virtual environments inhabited with autonomous characters whose behavior is driven by motion control assumptions is a promising approach to solve this problem. For example, an interesting problem in this field is navigation in an environment inhabited with other people. Typically, avoiding static obstacles together with other people displacing into the environment is a combinatory problem that strongly relies on the coupling between perception and action.

One of the main objectives of MimeTIC is to enhanche knowledge on human motion control by developing innovative experiments based on computer simulation and immersive environments. To this end, designing experimental protocols is a key point and some of the researchers in MimeTIC have developed this skill in biomechanics and perception-action coupling. Associating these researchers to experts in virtual human simulation, computational geometry and constraints solving enable us to contribute to enhance fundamental knowleged in human motion control.

3.2. Experiments in Virtual Reality

Understanding interaction between humans is very challenging because it addresses many complex phenomena including perception, decision-making, cognition and social behaviors. Moreover, all these phenomena are difficult to isolate in real situations, it is thus very complex to understand the influence of each of them on the interaction. It is then necessary to find an alternative solution that can standardize the experiments and that allows the modification of only one parameter at a time. Video was first used since the displayed experiment is perfectly repeatible and cut-offs (stop the video at a specific time before its end) allow having temporal information. Nevertheless, the absence of adapted viewpoint and stereoscopic vision does not provide depth information that are very meaningful. Moreover, during video recording session, the real human is acting in front of a camera and not an opponent. The interaction is then not a real interaction between humans.

Virtual Reality (VR) systems allow full standardization of the experimental situations and the complete control of the virtual environment. It is then possible to modify only one parameter at a time and observe its influence on the perception of the immersed subject. VR can then be used to understand what information are picked up to make a decision. Moreover, cut-offs can also be used to obtain temporal information about when these information are picked up. When the subject can moreover react as in real situation, his movement (captured in real time) provides information about his reactions to the modified parameter. Not only is the perception studied, but the complete perception-action loop. Perception and action are indeed coupled and influence each other as suggested by Gibson in 1979.

Finally, VR allows the validation of the virtual human models. Some models are indeed based on the interaction between the virtual character and the other humans, such as a walking model. In that case, there are two ways to validate it. First, they can be compared to real data (e.g. real trajectories of pedestrians). But such data are not always available and are difficult to get. The alternative solution is then to use VR. The validation of the realism of the model is then done by immersing a real subject in a virtual environment in which a virtual

character is controlled by the model. Its evaluation is then deduced from how the immersed subject reacts when interacting with the model and how realistic he feels the virtual character is.

3.3. Computational Geometry

Computational geometry is a branch of computer science devoted to the study of algorithms which can be stated in terms of geometry. It aims at studying algorithms for combinatorial, topological and metric problems concerning sets of points in Euclidian spaces. Combinatorial computational geometry focuses on three main problem classes: static problems, geometric query problems and dynamic problems.

In static problems, some input is given and the corresponding output needs to be constructed or found. Such problems include linear programming, Delaunay triangulations, and Euclidian shortest paths for instance. In geometric query problems, commonly known as geometric search problems, the input consists of two parts: the search space part and the query part, which varies over the problem instances. The search space typically needs to be preprocessed, in a way that multiple queries can be answered efficiently. Some typical problems are range searching, point location in a portioned space, nearest neighbor queries for instance. In dynamic problems, the goal is to find an efficient algorithm for finding a solution repeatedly after each incremental modification of the input data (addition, deletion or motion of input geometric elements). Algorithms for problems of this type typically involve dynamic data structures. Both of previous problem types can be converted into a dynamic problem, for instance, maintaining a Delaunay triangulation between moving points.

The Mimetic team works on problems such as crowd simulation, spatial analysis, path and motion planning in static and dynamic environments, camera planning with visibility constraints for instance. The core of those problems, by nature, relies on problems and techniques belonging to computational geometry. Proposed models pay attention to algorithms complexity to be compatible with performance constraints imposed by interactive applications.

MINT Project-Team

3. Research Program

3.1. Human-Computer Interaction

The scientific approach that we follow considers user interfaces as means, not an end: our focus is not on interfaces, but on interaction considered as a phenomenon between a person and a computing system [46]. We *observe* this phenomenon in order to understand it, i.e. *describe* it and possibly *explain* it, and we look for ways to significantly *improve* it. HCI borrows its methods from various disciplines, including Computer Science, Psychology, Ethnography and Design. Participatory design methods can help determine users' problems and needs and generate new ideas, for example [52]. Rapid and iterative prototyping techniques allow to decide between alternative solutions [47]. Controlled studies based on experimental or quasi-experimental designs can then be used to evaluate the chosen solutions [54]. One of the main difficulties of HCI research is the doubly changing nature of the studied phenomenon: people can both adapt to the system and at the same time adapt it for their own specific purposes [51]. As these purposes are usually difficult to anticipate, we regularly *create* new versions of the systems we develop to take into account new theoretical and empirical knowledge. We also seek to *integrate* this knowledge in theoretical frameworks and software tools to disseminate it.

3.2. Numerical and algorithmic real-time gesture analysis

Whatever is the interface, user provides some curves, defined over time, to the application. The curves constitute a gesture (positionnal information, yet may also include pressure). Depending on the hardware input, such a gesture may be either continuous (e.g. data-glove), or not (e.g. multi-touch screens). User gesture can be multi-variate (several fingers captured at the same time, combined into a single gesture, possibly involving two hands, maybe more in the context of co-located collaboration), that we would like, at higher-level, to be structured in time from simple elements in order to create specific command combinations. One of the scientific foundations of the research project is an algorithmic and numerical study of gesture, which we classify into three points:

- *clustering*, that takes into account intrinsic structure of gesture (multi-finger/multi-hand/multi-user aspects), as a lower-level treatment for further use of gesture by application;
- *recognition*, that identifies some semantic from gesture, that can be further used for application control (as command input). We consider in this topic multi-finger gestures, two-handed gestures, gesture for collaboration, on which very few has been done so far to our knowledge. On the contrary, in the case of single gesture case (i.e. one single point moving over time in a continuous manner), numerous studies have been proposed in the current literature, and interestingly, are of interest in several communities: HMM [55], Dynamic Time Warping [57] are well-known methods for computer-vision community, and hand-writing recognition. In the computer graphics community, statistical classification using geometric descriptors has previously been used [53]; in the Human-Computer interaction community, some simple (and easy to implement) methods have been proposed, that provide a very good compromise between technical complexity and practical efficiency [56].
- *mapping to application*, that studies how to link gesture inputs to application. This ranges from transfer function that is classically involved in pointing tasks [48], to the question to know how to link gesture analysis and recognition to the algorithmic of application content, with specific reference examples.

We ground our activity on the topic of numerical algorithm, expertise that has been previously achieved by team members in the physical simulation community (within which we think that aspects such as elastic deformation energies evaluation, simulation of rigid bodies composed of unstructured particles, constraint-based animation... will bring up interesting and novel insights within HCI community).

3.3. Design and control of haptic devices

Our scientific approach in the design and control of haptic devices is focused on the interaction forces between the user and the device. We search of controlling them, as precisely as possible. This leads to different designs compared to other systems which control the deformation instead. The research is carried out in three steps:

- *identification:* we measure the forces which occur during the exploration of a real object, for example a surface for tactile purposes. We then analyze the record to deduce the key components – *on user's point of view* – of the interaction forces.
- *design:* we propose new designs of haptic devices, based on our knowledge of the key components of the interaction forces. For example, coupling tactile and kinesthetic feedback is a promising design to achieve a good simulation of actual surfaces. Our goal is to find designs which leads to compact systems, and which can stand close to a computer in a desktop environment.
- *control:* we have to supply the device with the good electrical conditions to accurately output the good forces.

MORPHEO Project-Team

3. Research Program

3.1. Shape Acquisition

Multiple camera setups allow to acquire shapes, i.e. geometry, as well as their appearances, i.e. photometry, with a reasonable level of precision. However fundamental limitations still exist, in particular today's state-of-the-art approaches do not fully exploit the redundancy of information over temporal sequences of visual observations. Despite an increasing interest of the computer vision communities in the past years, the problem is still far from solved other than in specific situations with restrictive assumptions and configurations. Our goal in this research axis is to open the acquisition process to more general assumptions, e.g. no specific lighting or background conditions, scenes with evolving topologies, , and fully leverage temporal aspects of the acquisition process.

3.2. Bayesian Inference

Acquisition of 4D Models can often be conveniently formulated as a Bayesian estimation or learning problem. Various generative and graphical models can be proposed for the problems of occupancy estimation, 3D surface tracking in a time sequence, and motion segmentation. The idea of these generative models is to predict the noisy measurements (e.g. pixel values, measured 3D points or speed quantities) from a set of parameters describing the unobserved scene state, which in turn can be estimated using Bayes' rule to solve the inverse problem. The advantages of this type of modeling are numerous, as they enable to model the noisy relationships between observed and unknown quantities specific to the problem, deal with outliers, and allow to efficiently account for various types of priors about the scene and its semantics. Sensor models for different modalities can also easily be seamlessly integrated and jointly used, which remains central to our goals.

Since the acquisition problems often involve a large number of variables, a key challenge is to exhibit models which correctly account for the observed phenoma, while keeping reasonable estimation times, sometimes with a real-time objective. Maximum likelihood / maximum a posteriori estimation and approximate inference techniques, such as Expectation Maximization, Variational Bayesian inference, or Belief Propagation, are useful tools to keep the estimation tractable. While 3D acquisition has been extensively explored, the research community faces many open challenges in how to model and specify more efficient priors for 4D acquisition and temporal evolution.

3.3. Shape Analysis

Shape analysis has received much attention from the scientific community and recovering the intrinsic nature of shapes is currently an active research domain. Of particular interest is the study of human and animal shapes and their associated articulated underlying structures, i.e. skeletons, since applications are numerous, either in the entertainment industry or for medical applications, among others. Our main goals in this research axis are : the understanding of a shape's global structure, and a pose-independent classification of shapes.

3.4. Shape Tracking

Recovering the temporal evolution of a deformable surface is a fundamental task in computer vision, with a large variety of applications ranging from the motion capture of articulated shapes, such as human bodies, to the deformation of complex surfaces such as clothes. Methods that solve for this problem usually infer surface evolutions from motion or geometric cues. This information can be provided by motion capture systems or one of the numerous available static 3D acquisition modalities. In this inference, methods are faced with the challenging estimation of the time-consistent deformation of a surface from cues that can be sparse and noisy. Such an estimation is an ill posed problem that requires prior knowledge on the deformation to be introduced in order to limit the range of possible solutions. Our goal is to devise robust and accurate solutions based on new deformation models that fully exploit the geometric and photometric information available.

3.5. Motion Modeling

Multiple views systems can significantly change the paradigm of motion capture. Traditional motion capture systems provide 3D trajectories of a sparse set of markers fixed on the subject. These trajectories can be transformed into motion parameters on articulated limbs with the help of prior models of the skeletal structure. However, such skeletal models are mainly robotical abstractions that do not describe the true morphology and anatomical motions of humans and animals. On the other hand, 4D models (temporally consistent mesh sequences) provide dense motion information on body's shape while requiring less prior assumption. They represent therefore a new rich source of information on human and animal shape movements. The analysis of such data has nevertheless received few attention yet and tools still need to be developed which is our objective.

MULTISPEECH Team

3. Research Program

3.1. Introduction

As mentioned previously, MULTISPEECH is structured along three research directions that are associated to the three challenges previously described: explicit modeling of speech, statistical modeling of speech, and uncertainty in speech processing.

3.2. Explicit modeling of speech production and perception

Speech signals are the consequence of the deformation of the vocal tract under the effect of the movements of the jaw, lips, tongue, soft palate and larynx to modulate the excitation signal produced by the vocal cords or air turbulence. These deformations are visible on the face (lips, cheeks, jaw) through the coordination of different orofacial muscles and skin deformation induced by the latter. These deformations may also express different emotions. We should note that human speech expresses more than just phonetic content, to be able to communicate effectively. In this project, we address the different aspects related to speech production from the modeling of the vocal tract up to the production of audiovisual speech. On the one hand, we study the relationship from acoustic speech signal to vocal tract, in the context of acoustic-to-articulatory inversion, and from vocal tract to acoustic speech, in the context of articulatory synthesis. On the other hand, we work on expressive audiovisual speech synthesis, where both expressive acoustic speech and visual signals are generated from text. Phonetic contrasts used by the phonological system of any language result from constraints imposed by the nature of the human speech production apparatus. For a given language these contrasts are organized so as to guarantee that human listeners can identify sounds robustly. From the point of view of perception, these contrasts enable efficient processes of categorization in the peripheral and central human auditory system. The study of the categorization of sounds and prosody thus provides a complementary view on speech signals by focusing on the discrimination of sounds by humans, particularly in the context of language learning.

3.2.1. Articulatory modeling

Modeling speech production is a major issue in speech sciences. Acoustic simulation makes the link between articulatory and acoustic domains. Unfortunately this link cannot be fully exploited because there is almost always an acoustic mismatch between natural and synthetic speech generated with an articulatory model approximating the vocal tract. However, the respective effects of the geometric approximation, of the fact of neglecting some cavities in the simulation, of the imprecision of some physical constants and of the dimensionality of the acoustic simulation are still unknown. Hence, the first objective is to investigate the origin of the acoustic mismatch by designing more precise articulatory models, developing new methods to acquire tridimensional MRI data of the entire vocal tract together with denoised speech signals, and evaluating several approaches of acoustic simulation. This will enable the acoustic mismatch to be better controlled and the determination of the potential precision of inversion to be evaluated in particular.

Up to now, acoustic-to-articulatory inversion has been addressed as an instantaneous problem, articulatory gestures being recovered by concatenating local solutions via the determination of trajectories minimizing some articulatory cost. The second objective is thus to investigate how more elaborated strategies (a syllabus of primitive gestures, articulatory targets...) can be incorporated in the acoustic-to-articulatory inversion algorithms to take into account dynamic aspects.

This area of research relies on the equipment available in the laboratory to acquire articulatory data: articulograph Carstens AG501, head-neck antenna to acquire MRI of the vocal tract at Nancy Hospital, and multimodal acquisition system. Very few sites in France benefit from such a combination of acquisition devices.

3.2.2. Expressive acoustic-visual synthesis

Speech is considered as a bimodal communication means; the first modality is audio, provided by acoustic speech signals and the second one is visual, provided by the face of the speaker. Our research impacts both audiovisual and acoustic-only synthesis fields.

In our approach, the Acoustic-Visual Text-To-Speech synthesis (AV-TTS) is performed simultaneously with respect to its acoustic and visible components, by considering a bimodal signal comprising both acoustic and visual channels. A first AV-TTS system was developed resulting in a talking head; the system relied on 3D-visual data (3D markers on the face, data acquired by MAGRIT team) and on an extension of our non-uniform acoustic-unit concatenation text-to-speech synthesis system (SoJA). An important goal is to provide an audiovisual synthesis that is intelligible, both acoustically and visually. Thus, we continue working on adding visible components of the head through a tongue model where the tongue deformations come from EMA data analysis; and a lip-model to tackle the main recurrent problem of the lack of some lip markers in the 3D data. We will also improve the TTS engine to increase the accuracy of the unit selection simultaneously into the acoustic and visual domains (learning weights, feature selection...).

Another challenging research goal is to add expressivity in the AV-TTS. The expressivity comes through the acoustic signal (prosody aspects) and also through head and eyebrow movements. One objective is to add a prosodic component in the TTS engine in order to take into account some given prosodic entities such as emphasis, in order to highlight some important key words. Expressivity could be introduced before the unit selection step but also by developing algorithms intended to modify the parameters of prosody (in the acoustic domain, and in the visual domain as well). One intended approach will be to explore an expressivity measure at sound, syllable and/or sentence levels that describes the degree of perception or realization of an expression/emotion (audio and 3D domain). Such measures will be used as criteria in the selection process of the synthesis system. To tackle this issue we will also investigate Hidden Markov Model (HMM) based synthesis. The flexibility of the HMM-based approach enables the adjustment of the modeling parameters according to the available data and an easy adaption of the system to various conditions. This point will rely upon our experience in HMM modeling.

To acquire the facial data, we consider using marker-less motion capture system using a kinect-like system with a face tracking software. The software presents a user-friendly interface to track and visualize the motion in real time. Audio is also acquired synchronously with facial data. The advantage of this new system is to acquire rapidly the movements of the face with an acceptable quality. This system is used as an alternative relatively low-cost system to the VICON system.

3.2.3. Categorization of sounds and prosody for native and non-native speech

Discriminating speech sounds and prosodic patterns is the keystone of language learning whether in the mother tongue or in a second language. This issue is associated with the emergence of phonetic categories, i.e., classes of sounds which are related to phonemes, and prosodic patterns. The study of categorization is concerned not only with acoustic modeling but also with speech perception and phonology. Foreign language learning raises the issue of categorizing phonemes of the second language given the phonetic categories of the mother tongue. Thus, studies on the emergence of new categories, whether in the mother tongue (for people with language deficiencies) or in a second language, must rely upon studies on native and non-native acoustic realizations of speech sounds and prosody (i.e., at the segmental level and at the supra-segmental level). Moreover, as categorization is a perceptual process, studies on the emergence of categories must also rely on perceptual experiments.

Studies on native sounds have been an important research area of the team for years, leading to the notion of "selective" acoustic cues and the development of acoustic detectors. This know-how will be exploited in the study of non-native sounds. Concerning prosody, studies are focused on native and non-native realizations of modalities (e.g., question, affirmation, command ...), as well as non-native realizations of lexical accents and focus (emphasis). Results aim at providing automatic feedbacks to language learners with respect to acquisition of prosody as well as acquisition of a correct pronunciation of the sounds of the foreign language. Concerning the mother tongue we are interested in the monitoring of the process of sound categorization in the

long term (mainly at primary school) and its relation with the learning of reading and writing skills, especially for children with language deficiencies.

3.3. Statistical modeling of speech

Whereas the first research direction deals with the physical aspects of speech and its explicit modeling, this second research direction is concerned by investigating complex statistical models for speech data. Acoustic models are used to represent the pronunciation of the sounds or other acoustic events such as noises. Whether they are used for source separation, for speech recognition, for speech transcription, or for speech synthesis, the achieved performance strongly depends on the accuracy of these models, which is a critical aspect that is studied in the project. At the linguistic level, MULTISPEECH investigates models for handling the context (beyond the few preceding words currently handled by the n-gram models) and evolutive lexicons necessary when dealing with diachronic audio documents in order to overcome the limited size of the current static lexicons used, especially with respect to proper names. Statistical approaches are also useful for generating speech signals. Along this direction, MULTISPEECH mainly considers voice transformation techniques, with their application to pathological voices, and statistical speech synthesis applied to expressive multimodal speech synthesis.

3.3.1. Acoustic modeling

Acoustic modeling is a key issue for automatic speech recognition. Despite progress made for many years, acoustic modeling is still far from perfect, and current speech recognition applications rely on strong constraints (limited vocabulary, speaker adaptation, restricted syntax...) to achieve acceptable performance. As the acoustic models represent the acoustic realization of the sounds, they have to account for many variability sources, such as speaker characteristics, microphones, noises, etc. Extension of the HMM formalism based on the Dynamic Bayesian Networks (DBN) formalism are investigated further for handling such variability sources; as well as other approaches to dynamically constrain the search space according to known or estimated characteristics of the utterance being processed. Deep Neural Networks (DNN) based approaches will also be investigated as means of making speech recognition systems more accurate and robust. Speaker dependent modeling and speaker adaptation will also be investigated in relation with HMM-based speech synthesis and statistical voice conversion.

State-of-the-art speech recognition systems are still very sensitive to the quality of speech signals they have to deal with; their performance degrades rapidly when they deal with noisy signals. Accurate signal enhancement techniques are therefore essential to increase the robustness of both automatic speech recognition and speech-text alignment systems to noise and non-speech events. In MULTISPEECH, focus is set on Bayesian source separation techniques using multiple microphones and/or models of non-speech events. Some of the challenges include building a non-parametric model of the sources in the time-frequency-channel domain, linking the parameters of this model to the cepstral representation used in speech processing, modeling the temporal structure of environmental noise, and exploiting large audio data sets to automatically discover new models. Beyond the definition of such complex models, the difficulty is to design scalable estimation algorithms robust to overfitting, that will be integrated in the FASST [6] framework that was recently developed.

3.3.2. Linguistic modeling

MULTISPEECH investigates lexical and language models in speech recognition with a focus on improving the processing of proper names and the processing of spontaneous speech. Collaborations are ongoing with the SMarT team on linguistic modeling aspects.

Proper names are relevant keys in information indexing, but are a real problem in transcribing many diachronic spoken documents (such as radio or TV shows) which refer to data, especially proper names, that evolve over the time. This leads to the challenge of dynamically adjusting lexicons and language models through the use of the context of the documents or of some relevant external information possibly collected over the web. Random Indexing (RI) and Latent Dirichlet Allocation (LDA) are two possible approaches to be used for this purpose. Also, to overcome the limitations of current n-gram based language models, we investigate language

models defined on a continuous space in order to achieve a better generalization on unseen data, and to model long-term dependencies. This is achieved through neural network based approaches. We also want to introduce into these new models additional relevant information such as linguistic features, semantic relation, topic or user-dependent information.

Spontaneous speech utterances are often ill-formed and frequently contain disfluencies (hesitations, repetitions...) that degrade speech recognition performance. This is partly due to the fact that disfluencies are not properly represented in linguistic models estimated from clean text data (coming from newspapers for example); hence a particular effort will be set for improving the modeling of these events.

Attention will also be set on pronunciation lexicons in particular with respect to non-native speech and foreign names. Non-native pronunciation variants have to take into account frequent miss-pronunciations due to differences between mother tongue and target language phoneme inventories. Proper name pronunciation variants are a similar problem where difficulties are mainly observed for names of foreign origin that can be pronounced either in a French way or kept close to foreign origin native pronunciation. Automatic grapheme-to-phoneme state-of-the-art approaches, based for example on Joint Multigram Models (JMM) or Conditional Random Fields (CRF) will be further investigated and combined.

3.3.3. Speech generation by statistical methods

Voice conversion consists in building a function that transforms a given voice into another one. MULTISPEECH applies voice conversion techniques to enhance pathological voices that result from vocal folds problems, especially esophageal voice or pathological whispered voice. Voice conversion techniques are also of interest for text-to-speech synthesis systems as they aim at making possible the generation of new voice corpora (other kind of voice, or same voice with different kind of emotion).

In addition to the statistical aspects of the voice conversion approaches, signal processing is critical for good quality speech output. Information on the fundamental frequency is chaotic in the case of esophageal speech or non-existent in the case of the whispered voice. So after applying voice conversion techniques for enhancing pathological voices, the excitation spectrum must be predicted or corrected. That is the challenge that is addressed in the project. Also, in the context of acoustic feedback in foreign language learning, voice modification approaches (either statistical or not) will be investigated to modify the learner's (or teacher's) voice in order to emphasize the difference between the learner's acoustic realization and the expected realization.

Over the last few years statistical speech synthesis has emerged as an alternative to corpus-based speech synthesis. Speaker-dependent HMM modeling constitute the basis of such an approach. The announced advantages of the statistical speech synthesis are the possibility to deal with small amounts of speech resources and the flexibility for adapting models (for new emotions or new speaker), however, the quality is not as good as that of the concatenation-based speech synthesis. The reasons are twofold: first, parameters (F0, spectrum, duration...) are modeled independently and the models, even when taking into account dynamics, do not manage to generate parameters with a good precision. Second, the HMM generates sequences of feature vectors from which the actual speech signals are reconstructed, and this impacts on its quality. MULTISPEECH will focus on an hybrid approach, combining corpus-based synthesis, for its high-quality speech signal output, and HMM-based speech synthesis for its flexibility to drive selection, and the main challenge will be on its application to producing expressive audio-visual speech. One secondary objective will be to unify the HMM-based and the concatenation-based approaches.

3.4. Uncertainty estimation and exploitation in speech processing

After the explicit modeling presented and the statistical modeling that were previously described, we focus here on the uncertainty associated to some processing steps. Uncertainty stems from the high variability of speech signals and from imperfect models. For example, enhanced speech signals resulting from source separation are not exactly the clean original speech signals. Words or phonemes resulting from automatic speech recognition contain errors, and the phone boundaries resulting from automatic speech-text alignment

are not always correct, especially in acoustically degraded conditions. Hence it is important to know the reliability of the results and/or to estimate the uncertainty on the results.

3.4.1. Uncertainty and acoustic modeling

Because small distortions in the separated source signals can translate into large distortions in the cepstral features used for speech recognition, this limits the recognition performance on noisy data. One way to address this issue is to estimate the uncertainty on the separated sources in the form of their posterior distribution and to propagate this distribution, instead of a point estimate, through the subsequent feature extraction and speech decoding stages. Although major improvements have been demonstrated in proof-of-concept experiments using knowledge of the true uncertainty, accurate uncertainty estimation and propagation remains an open issue.

MULTISPEECH seeks to provide more accurate estimates of the posterior distribution of the separated source signals accounting for, e.g., posterior correlations over time and frequency which have not been considered so far. The framework of variational Bayesian (VB) inference appears to be a promising direction. Mappings learned on training data and fusion of multiple uncertainty estimators are also explored. The estimated uncertainties is then exploited for acoustic modeling in speech recognition and, in the future, also for speech-text alignment. This approach may later be extended to the estimation of the resulting uncertainty on the acoustic model parameters and the acoustic scores themselves.

3.4.2. Uncertainty and phonetic segmentation

The accuracy of the phonetic segmentation is important in several cases, as for example for the computation of prosodic features, for avoiding incorrect feedback to the learner in computer assisted foreign language learning, or for the post-synchronization of speech with face/lip images. Currently the phonetic boundaries obtained are quite correct on good quality speech, but the precision degrades significantly on noisy and non-native speech. Phonetic segmentation aspects will be investigated, both in speech recognition (i.e., spoken text unknown) and in forced alignment (i.e., when the spoken text is known). The first case (speech recognition) is connected with the computation of prosodic features for structuring speech recognition output, whereas the second case (forced alignment) is important in the context of non-native speech segmentation for automatic feedbacks in language learning.

In the same way that combining several speech recognition outputs leads to improved speech recognition performance, MULTISPEECH will investigate the combination of several speech-text alignments as a way of improving the quality of speech-text alignment and of determining which phonetic boundaries are reliable and which ones are not, and also for estimating the uncertainty on the boundaries. Knowing the reliability and/or the uncertainty on the boundaries will also be useful when segmenting speech corpora; this will help deciding which parts of the corpora need to be manually checked and corrected without an exhaustive checking of the whole corpus.

3.4.3. Uncertainty and prosody

Prosody information is also investigated as a means for structuring speech data (determining sentence boundaries, punctuation...) possibly in addition with syntactic dependencies (in collaboration with the SYNALP team). Structuring automatic transcription output is important for further exploitation of the transcription results such as easier reading after the addition of punctuation, or exploitation of full sentences in automatic translation. Prosody information is also necessary for determining the modality of the utterance (question or not), as well as determining accented words.

Prosody information comes from the fundamental frequency, the duration of the sounds and their energy. Any error in estimating these parameters may lead to a wrong decision. MULTISPEECH will investigate estimating the uncertainty on the duration of the phones (see uncertainty on phonetic boundaries above) and on the fundamental frequency, as well as how this uncertainty shall be propagated in the detection of prosodic phenomena such as accented words or utterance modality, or in the determination of the structure of the utterance. In a first approach, uncertainty estimation will rely on the comparison, and possibly the combination, of several estimators (several segmentation processes, several pitch algorithms).

OAK Project-Team

3. Research Program

3.1. Scalable and Expressive Techniques for the Semantic Web

The Semantic Web vision of a world-wide interconnected database of *facts*, describing *resources* by means of *semantics*, is coming within reach as the W3C's RDF (Resource Description Format) data model is gaining traction. The W3C Linking Open Data initiative has boosted the publication and interlinkage of a large number of datasets on the semantic web resulting to the Linked Open Data Cloud. These datasets of billions of RDF triples have been created and published online. Moreover, numerous datasets and vocabularies from different application domains are published nowadays as RDF graphs in order to facilitate community annotation and interlinkage of both scientific and scholarly data of interest. RDF storage, querying, and reasoning is now supported by a host of tools whose scalability and expressive power vary widely. Unsurprisingly, some of the most scalable tools draw upon the existing models and architecture for managing structured data. However, such tools often ignore the semantic aspects that make RDF interesting. For what concerns the semantics, a delicate balance must be found between expressive power and the efficiency of the resulting data management algorithms.

- The team works on identifying tractable dialects of RDF, amenable to highly efficient query answering algorithms, taking into account both data and semantics.
- Another line of research investigates the usage of RDF data and semantics to help structure, organize, and enrich structured documents from social media. Based on such a rich model, we devised novel query answering algorithms which attempt to explore efficiently the rich social dataset in order to return the most pertinent answers to the users, from a social, structured and semantic perspective. This research is related to the DIGICOSME LabEx grant "Structured, Social and Semantic Search".
- Last but not least, we investigate novel models and algorithms for efficient Semantic Web data management, going beyond the existing standard languages. We have finalized our proposal of an all-RDF data analytics framework, combining the rich structure and semantics of RDF with the power of analysis tools previously developed for relational data, such as analytical schemas and queries. Recent and ongoing work focuses on the automated selection of RDF analytical schemas as well as on efficient view-based analytical query answering strategies. The research is related to the "Investissement d'Avenir" project DATALYSE.

3.2. Massively Distributed Data Management Systems

Large and increasing data volumes have raised the need for distributed storage architectures. Among such architectures, computing in the cloud is an emerging paradigm massively adopted in many applications for the scalability, fault-tolerance and elasticity features it offers, which also allows for effortless deployment of distributed and parallel architectures. At the same time, interest in massively parallel processing has been renewed by the MapReduce model and many follow-up works, which aim at simplifying the deployment of massively parallel data management tasks in a cloud environment. For these reasons, cloud-based stores are an interesting avenue to explore for handling very large volumes of RDF data.

Our research aims at taking advantage of such widely available, large-scale distributed architectures to build scalable platforms for massively distributed management of complex data. We consider many different wide-scale distributed back-ends in this context, ranging from those provided by commercial cloud platforms to simple MapReduce and to more complex extensions thereof. In particular, we have considered the Stratosphere platform developed at TU Berlin, currently distributed by Apache under the name Flink.

This research is part of our participation to the Datalyse project previously mentioned, as well as the KIC EIT ICT Labs Europa activity, part of the "Computing in the Cloud" action line. We have completed our objectives within Europa, and our participation ended in 2014.

A recent development in this area is the start of our collaboration with social scientists from UNIV. PARIS-SUD, working on the management of innovation; we have started a collaborative research projects (ANR "Cloud-Based Organizational Design") where we perform an interdisciplinary analysis (both from a computing and from a business management perspective) on the adoption of cloud technologies within an enterprise.

3.3. Advanced Algorithms for Data Querying and Transformation

The *efficient* evaluation of queries over large databases remains a challenging task, to which certain optimization approaches based on static analysis of queries, data properties (such as integrity constraints), and indexing capabilities (such as materialized views) can still provide practically-relevant solutions. In this area, mainly for relational stores, we focus on query reformulation under constraints and views, as a uniform solution to problems such as view-based rewriting under constraints, semantic query optimization, and physical access path selection in query optimization.

With the increasing amount of available data, as well as the increasing complexity of data processing and transformations queries, for instance in applications such as relational data analysis or integration of Web data (e.g., XML or RDF), comes the need to better manage complex data transformations. In particular, it has become essential to analyze and debug data transformations. In this context, Oak has focused on verifying the semantic correctness of a declarative program that specifies a data transformation query, e.g., an SQL . In particular, we study one important sub-problem of data transformation analysis, namely the one of Why-Not questions. Such questions can explain to developers of complex data transformations or manipulations why their data transformation did not produce some specific results, although they expected them to do so.

3.4. Social Data Management and Crowdsourcing

The social Web blurs today the distinction between search, recommendation, and advertising (three paradigms for information access that have been so far considered mostly in separation). Our research in this area strives to find better adapted and scalable ways to answer information needs in the social Web, often by techniques at the intersection of databases, information retrieval, and data mining.

In particular, we study models and algorithms for personalized, or social-aware search in social applications. While progress has been made in this area, more remains to be done in order to address users' needs in practice, especially towards richer data models, and improving applicability and result relevance. For instance, when searching for tweets, their geographical location and recency may be as important for relevance as the textual and social aspects.

Furthermore, regarding quality of answers in response to searches, for various reasons (e.g., sparsity or tagging quality), meaningful results may often not be available. One response to this observation could be to turn to the crowd, the very users/publishers of the social media platform, and to turn this crowd into on-demand and query-driven sources of data. We study principled approaches for crowd selection (expert sourcing) and task assignment (data sourcing), in order to better answer ongoing social queries.

Beyond social links that represent just ties, a promising direction we also focus on in user-centric applications is to uncover implicit, potentially richer relationships from user interactions and to exploit them to improve core functionality such as search.

Moreover, we plan to investigate how crowdsourcing can be exploited to extract informations on user preferences, using techniques about noisy data management and provenance analysis.

ORPAILLEUR Project-Team

3. Research Program

3.1. From KDD to KDDK

Keywords: knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining

Knowledge discovery in databases is a process for extracting from large databases knowledge units that can be interpreted and reused. From an operational point of view, a KDD system includes databases, data mining modules, and interfaces for interactions, e.g. editing and visualization. The KDD process is based on three main operations: selection and preparation of the data, data mining, and finally interpretation of the extracted units.

The process of "knowledge discovery in databases guided by domain knowledge" extends the KDD cycle with a fourth step, where extracted units are represented within a knowledge base to be reused. The KDDK process –as implemented in the research work of the Orpailleur team– is based on *data mining methods* that are either symbolic or numerical:

- Symbolic methods are based on frequent itemsets search, association rule extraction, Formal Concept Analysis and extensions [113].
- Numerical methods are based on higher order stochastic models, namely second-order Hidden Markov Models (HMM2) and Hidden Markov fields (HMRF), which are especially designed for an efficient modeling of space and time [12].

The principle summarizing KDDK can be understood as a process going from complex data to knowledge units being guided by domain knowledge. Two original aspects can be underlined: (i) the knowledge discovery process is guided by domain knowledge at each step of the process, and (ii) the extracted units are embedded within knowledge-based systems for problem solving purposes.

One main operation in the research work of Orpailleur on KDDK is *classification*, which is a polymorphic process involved in modeling, mining, representing, and reasoning tasks. Moreover, the KDDK process is intended to feed knowledge-based systems working in application domains, e.g. agronomy, biology, chemistry, cooking and medicine, and also in the context of semantic web, text mining, information retrieval, and ontology engineering.

3.2. Knowledge Discovery guided by Domain Knowledge

Keywords: knowledge discovery, data mining, formal concept analysis, classification, frequent itemset search, association rule extraction, second-order Hidden Markov Models

Classification problems can be formalized by means of a class of objects (or individuals), a class of attributes (or properties), and a binary correspondence between the two classes, indicating for each individual-property pair whether the property applies to the individual or not. The properties may be features that are present or absent, or the values of a property that have been transformed into binary variables. Formal Concept Analysis (FCA) relies on the analysis of such binary tables and may be considered as a symbolic data mining technique to be used for extracting a set of formal concepts then organized within a concept lattice [113] (concept lattices are also known as "Galois lattices" [103]).

In parallel, the search for frequent itemsets and the extraction of association rules are well-known symbolic data mining methods, related to FCA (actually searching for frequent itemsets can be understood as traversing a concept lattice). Both processes usually produce a large number of items and rules, leading to the associated problems of "mining the sets of extracted items and rules". Some subsets of itemsets, e.g. frequent closed itemsets (FCIs), allow to find interesting subsets of association rules, e.g. informative association rules. This is why several algorithms are needed for mining data depending on specific applications [45].

Among useful patterns extracted from a database, frequent itemsets are usually thought to unfold "regularities" in the data, i.e. they are the witnesses of recurrent phenomena and they are consistent with the expectations of the domain experts. In some situations however, it may be interesting to search for "rare" itemsets, i.e. itemsets that do not occur frequently in the data (contrasting frequent itemsets). These correspond to unexpected phenomena, possibly contradicting beliefs in the domain. In this way, rare itemsets are related to "exceptions" and thus may convey information of high interest for experts in domains such as biology or medicine.

From the numerical point of view, a Hidden Markov Model (HMM2) is a stochastic process aimed at extracting and modeling a sequence of stationary distributions of events. Such models can be used for data mining purposes, especially for spatial and temporal data as they show good capabilities to locate patterns both in time and space domains.

Moreover, stochastic models have been designed to mine temporal sequences having a spatial dimension, for example the succession of land uses in a territory. One main Markovian assumption states that the temporal event succession in a given place depends only on the temporal event successions in neighboring points. By means of stochastic models such as hierarchical hidden Markov models and Markov random fields, it is possible to perform an unsupervised clustering of a spatial territory for discovering "patches" characterized by time and space regularities in their temporal successions.

3.3. Text Mining

Keywords: knowledge discovery form large collection of texts, text mining, information extraction, document annotation, ontologies

The objective of a text mining process is to extract useful knowledge units from large collections of texts [110]. The text mining process shows specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making text mining a particular task. To avoid information dispersion, a text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods.

From a KDDK perspective, text mining is aimed at extracting "interesting units" (nouns and relations) from texts with the help of domain knowledge encoded within an ontology (also useful for text annotation). Text mining is especially useful in the context of semantic web for ontology engineering [105]. In the Orpailleur team, the focus is put on the mining of real-world texts in application domains such as biology and medicine, using mainly symbolic data mining methods, and especially Formal Concept Analysis. Accordingly, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

3.4. Knowledge Systems and Semantic Web

Keywords: knowledge representation, ontology, description logics, classification-based reasoning, case-based reasoning, semantic web, information retrieval

Usually, people try to take advantage of the web by searching for information (navigation, exploration), and by querying documents using search engines (information retrieval). Then people try to analyze the obtained results, a task that may be difficult and tedious. Semantic web is an attempt for guiding search for information with the help of software agents, that are in charge of asking questions, searching for answers, classifying and interpreting the answers. However, a software agent may be able to read, understand, and manipulate information on the web, if and only if the knowledge necessary for achieving those tasks is available, and this is why ontologies are of main importance. Thus, there is a need for knowledge representation languages for annotating documents, describing the content of documents and giving a semantics to this content.

In particular, the knowledge representation language used for designing ontologies is the OWL language, which is based on description logics (DLs [100]). In OWL, knowledge units are represented within concepts (or classes), with attributes (properties of concepts, or relations, or roles), and individuals. The hierarchical organization of concepts (and relations) relies on a subsumption relation (i.e. a partial ordering).

The inference services are based on subsumption, concept and individual classification, two tasks related to "classification-based reasoning". Furthermore, classification-based reasoning can be extended into case-based reasoning (CBR), which relies on three main operations: retrieval, adaptation, and memorization. Given a target problem, retrieval consists in searching for a source (memorized) problem similar to the target problem. Then, the solution of the source problem is adapted to fulfill the constraints attached to the target problem, and possibly memorized for further reuse.

PANAMA Project-Team

3. Research Program

3.1. Axis 1: sparse models and representations

3.1.1. Efficient sparse models and dictionary design for large-scale data

Sparse models are at the core of many research domains where the large amount and high-dimensionality of digital data requires concise data descriptions for efficient information processing. Recent breakthroughs have demonstrated the ability of these models to provide concise descriptions of complex data collections, together with algorithms of provable performance and bounded complexity.

A crucial prerequisite for the success of today's methods is the knowledge of a "dictionary" characterizing how to concisely describe the data of interest. Choosing a dictionary is currently something of an "art", relying on expert knowledge and heuristics.

Pre-chosen dictionaries such as wavelets, curvelets or Gabor dictionaries, are based upon stylized signal models and benefit from fast transform algorithms, but they fail to fully describe the content of natural signals and their variability. They do not address the huge diversity underlying modern data much beyond time series and images: data defined on graphs (social networks, internet routing, brain connectivity), vector valued data (diffusion tensor imaging of the brain), multichannel or multi-stream data (audiovisual streams, surveillance networks, multimodal biomedical monitoring).

The alternative to a pre-chosen dictionary is a trained dictionary learned from signal instances. While such representations exhibit good performance on small-scale problems, they are currently limited to low dimensional signal processing due to the necessary training data, memory requirements and computational complexity. Whether designed or learned from a training corpus, dictionary-based sparse models and the associated methodology fail to scale up to the volume and resolution of modern digital data, for they intrinsically involve difficult linear inverse problems. To overcome this bottleneck, a new generation of efficient sparse models is needed, beyond dictionaries, which will encompass the ability to provide sparse and structured data representations as well as computational efficiency. For example, while dictionaries describe low-dimensional signal models in terms of their "synthesis" using few elementary building blocks called atoms, in "analysis" alternatives the low-dimensional structure of the signal is rather "carved out" by a set of equations satisfied by the signal. Linear as well as nonlinear models can be envisioned.

3.1.2. Compressive Learning

A flagship emerging application of sparsity is the paradigm of compressive sensing, which exploits sparse models at the analog and digital levels for the acquisition, compression and transmission of data using limited resources (fewer/less expensive sensors, limited energy consumption and transmission bandwidth, etc.). Besides sparsity, a key pillar of compressive sensing is the use of random low-dimensional projections. Through compressive sensing, random projections have shown their potential to allow drastic dimension reduction with controlled information loss, provided that the projected signal vector admits a sparse representation in some transformed domain. A related scientific domain, where sparsity has been recognized as a key enabling factor, is Machine Learning, where the overall goal is to design statistically founded principles and efficient algorithms in order to infer general properties of large data collections through the observation of a limited number of representative examples. Marrying sparsity and random low-dimensional projections with machine learning shall allow the development of techniques able to efficiently capture and process the information content of large data collections. The expected outcome is a dramatic increase of the impact of sparse models in machine learning, as well as an integrated framework from the signal level (signals and their acquisition) to the semantic level (information and its manipulation), and applications to data sizes and volumes of collections that cannot be handled by current technologies.

3.2. Axis 2: robust acoustic scene analysis

3.2.1. Compressive acquisition and processing of acoustic scenes

Acoustic imaging and scene analysis involve acquiring the information content from acoustic fields with a limited number of acoustic sensors. A full 3D+t field at CD quality and Nyquist spatial sampling represents roughly 10^6 microphones/m^3. Dealing with such high-dimensional data requires to drastically reduce the data flow by positioning appropriate sensors, and selecting from all spatial locations the few spots where acoustic sources are active. The main goal is to develop a theoretical and practical understanding of the conditions under which compressive acoustic sensing is both feasible and robust to inaccurate modeling, noisy measures, and partially failing or uncalibrated sensing devices, in various acoustic sensing scenarii. This requires the development of adequate algorithmic tools, numerical simulations, and experimental data in simple settings where hardware prototypes can be implemented.

3.2.2. Robust audio source separation

Audio signal separation consists in extracting the individual sound of different instruments or speakers that were mixed on a recording. It is now successfully addressed in the academic setting of linear instantaneous mixtures. Yet, real-life recordings, generally associated to reverberant environments, remain an unsolved difficult challenge, especially with many sources and few audio channels. Much of the difficulty comes from the combination of (i) complex source characteristics, (ii) sophisticated underlying mixing model and (iii) adverse recording environments. Moreover, as opposed to the "academic" blind source separation task, most applicative contexts and new interaction paradigms offer a variety of situations in which prior knowledge and adequate interfaces enable the design and the use of informed and/or manually assisted source separation methods.

The former METISS team has developed a generic and flexible probabilistic audio source separation framework that has the ability to combine various acoustic models such as spatial and spectral source models. A first objective is to instantiate and validate specific instances of this framework targeted to real-world industrial applications, such as 5.1 movie re-mastering, interactive music soloist control and outdoor speech enhancement. Extensions of the framework are needed to achieve real-time online processing, and advanced constraints or probabilistic priors for the sources at hand will be designed, while paying attention to computational scalability issues.

In parallel to these efforts, expected progress in sparse modeling for inverse problems shall bring new approaches to source separation and modeling, as well as to source localization, which is often an important first step in a source separation workflow. In particular, a research avenue consists in investigating physically motivated, lower-level source models, notably through sparse analysis of sound waves. This should be complementary with the modeling of non-point sources and sensors, and a widening of the notion of "source localization" to the case of extended sources (i.e., considering problems such as the identification of the directivity of the source as well as its spatial position), with a focus on boundary conditions identification. A general perspective is to investigate the relations between the physical structure of the source and the particular structures that can be discovered or enforced in the representations and models used for characterization, localization and separation.

3.3. Axis 3: large-scale audio content processing and self-organization

3.3.1. Motif discovery in audio data

Facing the ever-growing quantity of multimedia content, the topic of motif discovery and mining has become an emerging trend in multimedia data processing with the ultimate goal of developing weakly supervised paradigms for content-based analysis and indexing. In this context, speech, audio and music content, offers a particularly relevant information stream from which meaningful information can be extracted to create some form of "audio icons" (key-sounds, jingles, recurrent locutions, musical choruses, etc ...) without resorting to comprehensive inventories of expected patterns.

This challenge raises several fundamental questions that will be among our core preoccupations over the next few years. The first question is the deployment of motif discovery on a large scale, a task that requires extending audio motif discovery approaches to incorporate efficient time series pattern matching methods (fingerprinting, similarity search indexing algorithms, stochastic modeling, etc.). The second question is that of the use and interpretation of the motifs discovered. Linking motif discovery and symbolic learning techniques, exploiting motif discovery in machine learning are key research directions to enable the interpretation of recurring motifs.

On the application side, several use cases can be envisioned which will benefit from motif discovery deployed on a large scale. For example, in spoken content, word-like repeating fragments can be used for several spoken document-processing tasks such as language-independent topic segmentation or summarization. Recurring motifs can also be used for audio summarization of audio content. More fundamentally, motif discovery paves the way for a shift from supervised learning approaches for content description to unsupervised paradigms where concepts emerge from the data.

3.3.2. Structure modeling and inference in audio and musical contents

Structuring information is a key step for the efficient description and learning of all types of contents, and in particular audio and musical contents. Indeed, structure modeling and inference can be understood as the task of detecting dependencies (and thus establishing relationships) between different fragments, parts or sections of information content.

A stake of structure modeling is to enable more robust descriptions of the properties of the content and better model generalization abilities that can be inferred from a particular content, for instance via cache models, trigger models or more general graphical models designed to render the information gained from structural inference. Moreover, the structure itself can become a robust descriptor of the content, which is likely to be more resistant than surface information to a number of operations such as transmission, transduction, copyright infringement or illegal use.

In this context, information theory concepts will be investigated to provide criteria and paradigms for detecting and modeling structural properties of audio contents, covering potentially a wide range of application domains in speech content mining, music modeling or audio scene monitoring.

PERCEPTION Project-Team

3. Research Program

3.1. Audio-Visual Scene Analysis

From 2006 to 2009, R. Horaud was the scientific coordinator of the collaborative European project POP (Perception on Purpose), an interdisciplinary effort to understand visual and auditory perception at the crossroads of several disciplines (computational and biological vision, computational auditory analysis, robotics, and psychophysics). This allowed the PERCEPTION team to launch an interdisciplinary research agenda that has been very active for the last five years. There are very few teams in the world that gather scientific competences spanning computer vision, audio signal processing, machine learning and human-robot interaction. The fusion of several sensorial modalities resides at the heart of the most recent biological theories of perception. Nevertheless, multi-sensor processing is still poorly understood from a computational point of view. In particular and so far, audio-visual fusion has been investigated in the framework of speech processing using close-distance cameras and microphones. The vast majority of these approaches attempt to model the temporal correlation between the auditory signals and the dynamics of lip and facial movements. Our original contribution has been to consider that audio-visual localization and recognition are equally important. We have proposed to take into account the fact that the audio-visual objects of interest live in a three-dimensional physical space and hence we contributed to the emergence of *audio-visual scene analysis* as a scientific topic in its own right. We proposed several novel statistical approaches based on supervised and unsupervised mixture models. The *conjugate mixture model* (CMM) is an unsupervised probabilistic model that allows to cluster observations from different modalities (e.g., vision and audio) living in different mathematical spaces [9], [16]. We thoroughly investigated CMM, provided practical resolution algorithms and studied their convergence properties. We developed several methods for sound localization using two or more microphones [15]. The *Gaussian locally-linear model* (GLLiM) is a partially supervised mixture model that allows to map high-dimensional observations (audio, visual, or concatenations of audio-visual vectors) onto low-dimensional manifolds with a partially known structure [19]. This model is particularly well suited for perception because it encodes both observable and unobservable phenomena. A variant of this model, namely *probabilistic piecewise affine mapping* has also been proposed and successfully applied to the problem of sound-source localization and separation [18]. The European project HUMAVIPS (2010-2013), coordinated by R. Horaud, applied audio-visual scene analysis to human-robot interaction.

3.2. Binocular Vision

Stereoscopy is one of the most studied topics in biological and computer vision. Nevertheless, classical approaches of addressing this problem fail to integrate eye/camera vergence. From a geometric point of view, the integration of vergence is difficult because one has to re-estimate the epipolar geometry at every new eye/camera rotation. From an algorithmic point of view, it is not clear how to combine depth maps obtained with different eyes/cameras relative orientations. Therefore, we addressed the more general problem of binocular vision that combines the low-level eye/camera geometry, sensor rotations, and practical algorithms based on global optimization [4], [11]. We studied the link between mathematical and computational approaches to stereo (global optimization and Markov random fields) and the brain plausibility of some of these approaches: indeed, we proposed an original mathematical model for the complex cells in visual-cortex areas V1 and V2 that is based on steering Gaussian filters and that admits simple solutions [5]. This addresses the fundamental issue of how local image structure is represented in the brain/computer and how this structure is used for estimating a dense disparity field. Therefore, the main originality of our work is to address both computational and biological issues within a unifying model of binocular vision. Another equally important problem that still remains to be solved is how to integrate binocular depth maps over time. Recently, we have addressed this problem and proposed a semi-global optimization framework that starts with sparse yet reliable matches and proceeds with propagating them over both space and time. The concept of seed-match propagation has then been extended to TOF-stereo fusion.

3.3. Binaural Hearing

Audio-visual fusion algorithms necessitate that the two modalities are represented in the same mathematical space. Binaural hearing allows to extract sound-source localization (SSL) information from the acoustic signals recorded with two microphones. We have developed several methods, that perform sound localization in the temporal and the spectral domains. If a direct path is assumed, one can exploit the *time difference of arrival* (TDOA) between two microphones to recover the position of the sound source with respect to the position of the two microphones. The solution is not unique in this case, the sound source lies onto a 2D manifold. However, if one further assumes that the sound source lies in a horizontal plane, it is then possible to extract the azimuth. We used this approach to predict possible sound locations in order to estimate the direction of a speaker [16]. We also developed a geometric formulation and we showed that with four non-coplanar microphones the azimuth and elevation of a single source can be estimated without ambiguity [15]. We also investigated SSL in the spectral domain. This exploits the filtering effects of the head related transfer function (HRTF): there is a different HRTF for the left and right microphones. The interaural spectral features, namely the ILD (interaural level difference) and IPD (interaural phase difference) can be extracted from the short-time Fourier transforms of the two signals. The sound direction is encoded in these interaural features but it is not clear how to make SSL explicit in this case. We proposed a supervised learning formulation that estimates a mapping from interaural spectral features (ILD and IPD) to source directions using two different setups: audio-motor learning [18] and audio-visual learning [24]. Currently we generalize this approach to an arbitrary number of microphones.

3.4. Visual Reconstruction With Multiple Color and Depth Cameras

For the last decade, one of the most active topics in computer vision has been the visual reconstruction of objects, people, and complex scenes using a multiple-camera setup. The PERCEPTION team has pioneered this field and by 2006 several team members published seminal papers in the field. Recent work has concentrated onto the robustness of the 3D reconstructed data using probabilistic outlier rejection techniques combined with algebraic geometry principles and linear algebra solvers [14]. Subsequently, we proposed to combine 3D representations of shape (meshes) with photometric data [12]. The originality of this work was to represent photometric information as a scalar function over a discrete Riemannian manifold, thus *generalizing image analysis to mesh and graph analysis*. Manifold equivalents of local-structure detectors and descriptors were developed [13]. The outcome of this pioneering work has been twofold: the formulation of a new research topic now addressed by several teams in the world, and allowed us to start a three year collaboration with Samsung Electronics. We developed the novel concept of *mixed camera systems* combining high-resolution color cameras with low-resolution depth cameras [34], [21], [20]. Together with our start-up company 4D Views Solutions and with Samsung, we developed the first practical depth-color multiple-camera multiple-PC system and the first algorithms to reconstruct high-quality 3D content.

3.5. Registration, Tracking and Recognition of People and Actions

The analysis of articulated shapes has challenged standard computer vision algorithms for a long time. There are two difficulties associated with this problem, namely how to represent articulated shapes and how to devise robust registration and tracking methods. We addressed both these difficulties and we proposed a novel kinematic representation that integrates concepts from robotics and from the geometry of vision. In 2008 we proposed a method that parameterizes the occluding contours of a shape with its intrinsic kinematic parameters, such that there is a direct mapping between observed image features and joint parameters [10]. This deterministic model has been motivated by the use of 3D data gathered with multiple cameras. However, this method was not robust to various data flaws and could not achieve state-of-the-art results on standard dataset. Subsequently, we addressed the problem using probabilistic generative models. We formulated the problem of articulated-pose estimation as a maximum-likelihood with missing data and we devised several tractable algorithms [8], [7]. We proposed several expectation-maximization procedures applied to various articulated shapes: human bodies, hands, etc. In parallel, we proposed to segment and register articulated shapes represented with graphs by embedding these graphs using the spectral properties of graph Laplacians

[17]. This turned out to be a very original approach that has been followed by many other researchers in computer vision and computer graphics.

POTIOC Project-Team

3. Research Program

3.1. Introduction

The project of team potioc is oriented along three axes:

- Understanding humans interacting with the digital world
- Creating interactive systems
- Exploring new applications and usages

These axes are depicted in Figure 2 .

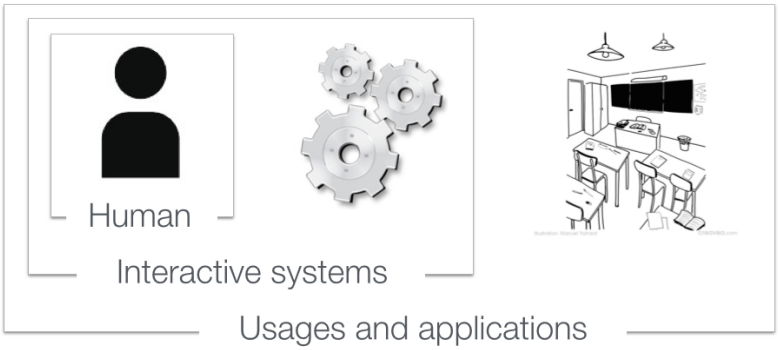


Figure 2. The three axes of the potioc team objectives.

Objective 1 is centered on the human sensori-motor and cognitive abilities, as well as user strategies and preferences, for completing interaction tasks. Our contributions for this objective are a better understanding of humans interacting with interactive systems. The impact of this objective is mainly at a fundamental level.
In objective 2, our goal is to create interactive systems. This may include hardware parts where new input and output modalities are explored. This also includes software parts, that are strongly linked to the underlying hardware components. Our contribution in objective 2 is to develop (hardware/software) interaction techniques allowing humans to perform interaction tasks.
Finally, in objective 3, we consider interaction at a higher level, taking into account factors that are linked to specifc application domains and usages. Our contribution in this area is the exploration and the emergence of new applications and usages that take benefit from the developments of the project. With this objective, we target mainly a societal impact.
Of course, strong links exist between the three objectives of the project. For example, the results obtained in objective 1 guide the development of objective 2. Inversely, new systems developed in objective 2 may feed research questions of objective 1. There exists similar links with objective 3.

3.2. Objective 1: Understanding humans interacting with the digital world

Our first objective is centered on the human side. Our finality is not to enhance the general knowledge about the human being as a research team in psychology would do. Instead, we focus on human skills and behaviors during interaction processes. To this end, we conduct experiments that allow us to better understand what the users like, where and why they have difficulties. Thanks to these investigations, we are able to design interaction techniques and systems (described in Objective 2) that are well suited to the targeted users. We believe that this fundamental piece of work is the first step that is required for the design of usable popular interactions. We are particularly interested in 3D interaction tasks for which we design dedicated experiments. We also propose a new approach based on physiological and brain (ElectroEncephaloGraphy - EEG) signals for the evaluation of these interactions.

3.2.1. Interacting with 3D

In the scope of the national project InSTInCT (ANR), we have studied how users tend to interact with a touchscreen for interacting with 3D content. Indeed, whereas such kind of interaction has been extensively studied for 2D contexts, it has been little explored in 3D. However, we believe that it is fundamental to understand users' strategies and preferences well in order to promote 3D interaction on touch screens. We conducted a set of experiments to investigate such kind of interaction. We proposed guidelines to help designers in the creation of more user friendly tools. Such kind of study led to the design of tBox. We also conducted experiments to better understand how users manage to control finger pressure, and how they tend to use this input modality. In another work, we have studied the impact of directness when manipulating 3D content on multitouch screens. This allowed us to gain knowledge about users performance in touch-based interaction.

3.2.2. Evaluating 3DUIs with physiological signals

We recently started to explore a new approach to HCI evaluations: using various physiological signals, and notably EEG signals, as a new complementary tool to assess objectively and more precisely the ergonomic quality of a given 3DUI. In particular we aim at using physiological signals to identify where and when the pros and cons of this interface are, based on the user's mental state during interaction. For instance, estimating the user's mental workload during interaction can give insights about where and when the interface is cognitively difficult to use. Such tools could prove very promising to improve evaluations by complementing existing tools (e.g., questionnaires or interviews) that can suffer from reporting bias, can disturb the user, or only provide an a-posteriori global (but undetailled) evaluation of the interaction. So far, we studied the different kinds of mental states that can be estimated from EEG signals and that are valuable for HCI and user evaluations. We also obtained promising first results suggesting that the level of comfort during stereoscopic visualization could be estimated from EEG signals, hence opening the way to faster, more objective and more individualized stereoscopic display design and calibration. Still with the objective of estimating various users' mental states to refine system evaluations and users' understanding, we explored mental stress (a.k.a., mental workload) and social stress (pressure due to a social evaluation) estimation from brain and physiological signals. To this end, we first had to design a protocol to induce mental stress and social stress, which we did successfully. Then, we were able to calibrate stress recognition from EEG and physiological signals as well as to assess the accuracy of the stress estimators. Then, we managed to robustly estimate mental stress levels from EEG and physiological signals (EEG being the most robust modality), even accross different contexts, here accross different levels of social stress. This is an interesting step towards robust estimation of mental stress in realistic conditions. Finally, we also studied and reviewed emotion recognition from EEG signals, which, again, is another interesting mental state to consider during an HCI evaluation.

3.2.3. Interacting with Brain-Computer Interfaces

Finally, we also studied how humans interact with a specific HCI: Brain-Computer Interfaces (BCI). Indeed, although EEG-based BCIs are very promising for numerous applications, e.g., rehabilitation or gaming, they mostly remain prototypes not used outside laboratories, due to their low reliability. Poor BCI performances are partly due to imperfect EEG signal processing algorithms but also to the user, who may not be able to

produce reliable EEG patterns. Indeed, BCI use is a skill, requiring the user to be properly trained to achieve BCI control. If he/she cannot perform the desired mental commands, no signal processing algorithm could identify them. Therefore, rather than improving EEG signal processing alone (which is what most current BCI research is about), we proposed to also guide users to learn BCI control mastery. We actually studied some theoretical models and guidelines from psychology and cognitive sciences about human learning, which revealed the many theoretical limitations of current standard BCI training approaches. We also conducted some actual experiments to further illustrate some limitations of current BCI training protocols and try to understand and analyse them. Finally, we explored new feedback types and new EEG visualization techniques in order to help users to learn BCI control skills more efficiently. These new feedback and visualizations notably aim at providing BCI users with more information about their EEG patterns using, in order to identify more easily relevant BCI control strategies, as well as motivating and engaging them in the learning task. This was achieved using augmented reality displays of the activity on the whole cortex - using an approach entitled the "Mind-Mirror", or by using multiplayer video game-based BCI training. Overall, this line of research seem largely unexplored but promising, and we are currently investing increasingly more research efforts into it.

3.3. Objective 2: Creating interactive systems

Our objective here is to create interactive systems and design interaction techniques dedicated to the completion of interaction tasks. We divide our work into three main categories:

- Interaction techniques based on existing Input/Output (IO) devices.
- New IO and related techniques.
- BCI and physiological computing.

3.3.1. Interaction techniques based on existing Input/Output (IO) devices

When using desktop IO (i.e., based on mouse/keyboards/monitors), a big challenge is to design interaction techniques that allow users to complete 3D interaction tasks. Indeed, the desktop IO space that is mainly dedicated to the completion of 2D interaction task is not well suited to 3D content and, consequently, 3D user interfaces need to be designed with a great care. We have proposed a state of the art that describes the major approaches and techniques in this area. In the past few years, we have been particularly interested in the problem of interaction when the 3D content is displayed on a touchscreen. Indeed, standard (2D) HCI has evolved from mouse to touch input, and numerous research projects have been conducted. At the opposite, in 3D, very little work has been proposed. We have contributed to move desktop 3D UIs from the mouse to the touch paradigm; what we used to do with mice in front of a screen does not work well on touch devices anymore. To face this problem, we have focused on touch-based 3D UIs. The first work brought tBox , a new 3D transformation widget designed for manipulating 3D objects on touch screens. In a second work, we have explored several strategies for navigating in 3D digital cities from touch inputs in collaboration with our industrial partners Vectuel and Mappy/PagesJaunes.

3.3.2. New IO and related techniques

In Potioc, we are interested in exploring new IO modalities that may make interaction easier, more engaging and motivating. In the past few years, we have designed new interactive systems that exploit unconventional IO modalities. Stereoscopic visualization has a great potential for the understanding of 3D content. On the other hand, interaction with such stereoscopic environments is generally diffcult. To face this problem, we have conceived Toucheo, a new system that exploits stereoscopic visualization and touch input. We have also contributed to the design of a system that exploits 3D spatial and touch input in a stereoscopic 3D environment. In the scope of immersive VR, we have also proposed some extensions of the current IO space. In particular, we presented a new input device that has been specifically designed to play music in an immersive VR environment. It mixes graphical and percussion based interaction. Another example is the SIMCA project where we have build a gateway simulator composed of numerous screens, video projectors and tracking systems. Tangible interaction has also been a subject of interest for us. Indeed, we believe that manipulating directly physical objects for interacting with the digital world has a great potential, in particular when the

general public is targeted. In this direction, we have notably proposed PapARt, a system that mixes physical drawing and augmented reality. With this system, the computer disappears, and the user interacts with the digital content as he or she would do with physical content. Another example is Rouages where musicians play with physical midi instruments that are augmented with virtual information to provide rich experiences to the audience. Our more recent contribution is Teegi, a new system based on a unique combination of spatial augmented reality, tangible interaction and real-time neurotechnologies. With Teegi, a user can visualize and analyze his or her own brain activity in real-time, on a tangible character that can be easily manipulated, and with which it is possible to interact.

3.3.3. BCI and physiological computing

As part of our research on the design of interactive systems based on physiological signals, and in particular brain signals (for BCI design) we conducted a number of research projects on EEG signal processing and classification. Indeed, in order to design practical BCI that can be used outside the lab, there is a need for robust EEG signal processing algorithm with the long-term objective to correctly recognise the users' mental commands (and thus EEG patterns) anytime and anywhere. To do so, we first explored and designed new features to represent EEG signals. We notably explored multifractal cumulants and predictive complexity features, waveform length features with an optimal spatial filter that we designed, as well as phase-locking value features (i.e., functional connectivities between brain areas), also with an optimal spatial filter we designed. All such features proved useful to classify EEG signals, and, more importantly, increased BCI classification performances (by 2 to 4% on average) when combined with the gold standard features, namely, band power features. To make BCI more robust to noise and non-stationarities, we proposed to integrate a-priori knowledge into machine learning algorithms. Such knowledge represents any information we have about what should be a good filter or classifier for instance. We successfully demonstrated this approach to learn robust and stable spatial filters. Finally, we worked on reducing the long and tedious BCI calibration times, by making the design of a BCI possible from very few training EEG signals. To do so, we proposed to generate artificial EEG signals from the few EEG trials initially available, in order to augment the training set size in a relevant way. This enabled us to calibrate BCI systems with 2 to 3 times less data than standard designs, while maintaining similar classification performances, hence effectively reducing the calibration time by 2 or 3.

3.4. Objective 3: Exploring new applications and usages

Objective 3 is centered on the applications and usages. Beyond the human sensori-motor and cognitive skills (Objective 1), and the hardware and software components (Objective 2), Objectives 3 takes into account broader criteria for the emergence of new usages and applications in various areas, and in particular in the scope of learning, popularization of science, art and entertainment. Our goal here is not to develop full-packaged end-user applications. Instead, our contribution is to stimulate the evolution of current applications with new engaging interactive systems.

3.4.1. Popularization of science

In the scope of popularization of science, we have built a strong partnership with Cap Sciences, which is a center dedicated to the popularization of science in Bordeaux that is visited by thousands of visitors every month. This was initiated with the ANR national project InSTInCT, whose goal was to study the benefits of 3D touch-based interaction in public exhibitions. This project has led to the creation of a Living Lab where several systems developed by Potioc are tested by the visitors. This provides us with interesting feedback that goes beyond the feedback we can obtain in our controlled lab-experiments. In the scope of archeology, we also contributed to a new system dedicated to public exhibitions, and we collected the current work around the world in this area in a dedicated special issue of a journal. We also contributed to an experiment at "Palais de la découverte" in Paris, where hundreds of visitors have experimented with PapARt (Figure 3).

Figure 3. PapARt used as a mediation tool at "Palais de la découverte"

3.4.2. Education

In the scope of education, we are currently collaborating with Stṕhanie Fleck from Université de Lorraine for exploring new interactive systems that enhance learning processes. Furthermore, we have launched a project with colleagues in the scope of teaching optical phenomena in optics. Our project HOBIT aims at developing an Hybrid Optical Bench for Innovative Teaching.

3.4.3. Art

In the scope of Art, we are convinced that the work that is conducted in Potioc may benefit to creation from the artist point of view, and it may open new interactive experiences from the audience point of view. We have conducted work with colleagues who are specialists in digital music, and with musicians. This led to several scientific publications and live artistic performances. We have also worked with an architect in order to explore neurodesign, i.e., the use of neural signals for design, here for the design of artistic shapes. Furthermore, we continued exploring the artistic domain in the scope of interactive juggling.

3.4.4. Entertainment

In the scope of entertainement, we notably explored BCI-based gaming and non-medical applications of BCI. In particular, we studied and analyzed how BCI could be used as a control channel for virtual reality and gaming applications, as well as the pros and cons of BCI-based gaming. We also proposed and studied a multiplayer BCI-based game. Our work so far suggests that BCI-based gaming and virtual reality applications are feasible and promising, but that many research challenges are still to be overcome for widespread use. In another example in the field of entertainment we studied several input modalities for playing a game in mobile AR.

PRIMA Project-Team

3. Research Program

3.1. Situation Models for Context Aware Systems and Services

Context Awareness, Smart Spaces

3.1.1. Summary

Over the last few years, the PRIMA group has pioneered the use of context aware observation of human activity in order to provide non-disruptive services. In particular, we have developed a conceptual framework for observing and modeling human activity, including human-to-human interaction, in terms of situations.

Encoding activity in situation models provides a formal representation for building systems that observe and understand human activity. Such models provide scripts of activities that tell a system what actions to expect from each individual and the appropriate behavior for the system. A situation model acts as a non-linear script for interpreting the current actions of humans, and predicting the corresponding appropriate and inappropriate actions for services. This framework organizes the observation of interaction using a hierarchy of concepts: scenario, situation, role, action and entity. Situations are organized into networks, with transition probabilities, so that possible next situations may be predicted from the current situation.

Current technology allows us to handcraft real-time systems for a specific services. The current hard challenge is to create a technology to automatically learn and adapt situation models with minimal or no disruption of human activity. An important current problem for the PRIMA group is the adaptation of Machine Learning techniques for learning situation models for describing the context of human activity.

3.1.2. Detailed Description

Context Aware Systems and Services require a model for how humans think and interact with each other and their environment. Relevant theories may be found in the field of cognitive science. Since the 1980's, Philippe Johnson-Laird and his colleagues have developed an extensive theoretical framework for human mental models [45], [46]. Johnson Laird's "situation models", provide a simple and elegant framework for predicting and explaining human abilities for spatial reasoning, game playing strategies, understanding spoken narration, understanding text and literature, social interaction and controlling behavior. While these theories are primarily used to provide models of human cognitive abilities, they are easily implemented in programmable systems [34], [33].

In Johnson-Laird's Situation Models, a situation is defined as a configuration of relations over entities. Relations are formalized as N-ary predicates such as beside or above. Entities are objects, actors, or phenomena that can be reliably observed by a perceptual system. Situation models provide a structure for organizing assemblies of entities and relations into a network of situations. For cognitive scientists, such models provide a tool to explain and predict the abilities and limitations of human perception. For machine perception systems, situation models provide the foundation for assimilation, prediction and control of perception. A situation model identifies the entities and relations that are relevant to a context, allowing the perception system to focus limited computing and sensing resources. The situation model can provide default information about the identities of entities and the configuration of relations, allowing a system to continue to operate when perception systems fail or become unreliable. The network of situations provides a mechanism to predict possible changes in entities or their relations. Finally, the situation model provides an interface between perception and human centered systems and services. On the one hand, changes in situations can provide events that drive service behavior. At the same time, the situation model can provide a default description of the environment that allows human-centered services to operate asynchronously from perceptual systems.

We have developed situation models based on the notion of a script. A theatrical script provides more than dialog for actors. A script establishes abstract characters that provide actors with a space of activity for expression of emotion. It establishes a scene within which directors can layout a stage and place characters. Situation models are based on the same principle.

A script describes an activity in terms of a scene occupied by a set of actors and props. Each actor plays a role, thus defining a set of actions, including dialog, movement and emotional expressions. An audience understands the theatrical play by recognizing the roles played by characters. In a similar manner, a user service uses the situation model to understand the actions of users. However, a theatrical script is organised as a linear sequence of scenes, while human activity involves alternatives. In our approach, the situation model is not a linear sequence, but a network of possible situations, modeled as a directed graph.

Situation models are defined using roles and relations. A role is an abstract agent or object that enables an action or activity. Entities are bound to roles based on an acceptance test. This acceptance test can be seen as a form of discriminative recognition.

There is no generic algorithm capable of robustly recognizing situations from perceptual events coming from sensors. Various approaches have been explored and evaluated. Their performance is very problem and environment dependent. In order to be able to use several approaches inside the same application, it is necessary to clearly separate the specification of scenario and the implementation of the program that recognizes it, using a Model Driven Engineering approach. The transformation between a specification and its implementation must be as automatic as possible. We have explored three implementation models :

- *Synchronized petri net*. The Petri Net structure implements the temporal constraints of the initial context model (Allen operators). The synchronisation controls the Petri Net evolution based on roles and relations perception. This approach has been used for the Context Aware Video Acquisition application.
- *Fuzzy Petri Nets*. The Fuzzy Petri Net naturally expresses the smooth changes of activity states (situations) from one state to another with gradual and continuous membership function. Each fuzzy situation recognition is interpreted as a new proof of the recognition of the corresponding context. Proofs are then combined using fuzzy integrals. This approach has been used to label videos with a set of predefined scenarios (context).
- *Hidden Markov Model*. This probabilistic implementation of the situation model integrates uncertainty values that can both refer to confidence values for events and to a less rigid representation of situations and situations transitions. This approach has been used to detect interaction groups and to determinate who is interacting with whom and thus which interaction groups are formed.

Currently situation models are constructed by hand. Our challenge is to provide a technology by which situation models may be adapted and extended by explicit and implicit interaction with the user. An important aspect of taking services to the real world is an ability to adapt and extend service behaviour to accommodate individual preferences and interaction styles. Our approach is to adapt and extend an explicit model of user activity. While such adaptation requires feedback from users, it must avoid or at least minimize disruption. We are currently exploring reinforcement learning approaches to solve this problem.

With a reinforcement learning approach, the system is rewarded and punished by user reactions to system behaviours. A simplified stereotypic interaction model assures a initial behaviour. This prototypical model is adapted to each particular user in a way that maximizes its satisfaction. To minimize distraction, we are using an indirect reinforcement learning approach, in which user actions and consequences are logged, and this log is periodically used for off-line reinforcement learning to adapt and refine the context model.

Adaptations to the context model can result in changes in system behaviour. If unexpected, such changes may be disturbing for the end users. To keep user's confidence, the learned system must be able to explain its actions. We are currently exploring methods that would allow a system to explain its model of interaction. Such explanation is made possible by explicit describing context using situation models.

The PRIMA group has refined its approach to context aware observation in the development of a process for real time production of a synchronized audio-visual stream based using multiple cameras, microphones and other information sources to observe meetings and lectures. This "context aware video acquisition system" is an automatic recording system that encompasses the roles of both the cameraman and the director. The system determines the target for each camera, and selects the most appropriate camera and microphone to record the current activity at each instant of time. Determining the most appropriate camera and microphone requires a model of activities of the actors, and an understanding of the video composition rules. The model of the activities of the actors is provided by a "situation model" as described above.

In collaboration with France Telecom, we have adapted this technology to observing social activity in domestic environments. Our goal is to demonstrate new forms of services for assisted living to provide non-intrusive access to care as well to enhance informal contact with friends and family.

3.2. Service Oriented Architectures for Intelligent Environments

Software Architecture, Service Oriented Computing, Service Composition, Service Factories, Semantic Description of Functionalities

Intelligent environments are at the confluence of multiple domains of expertise. Experimenting within intelligent environments requires combining techniques for robust, autonomous perception with methods for modeling and recognition of human activity within an inherently dynamic environment. Major software engineering and architecture challenges include accomodation of a heterogeneous of devices and software, and dynamically adapting to changes human activity as well as operating conditions.

The PRIMA project explores software architectures that allow systems to be adapt to individual user preferences. Interoperability and reuse of system components is fundamental for such systems. Adopting a shared, common Service Oriented Architecture (SOA) architecture has allowed specialists from a variety of subfields to work together to build novel forms of systems and services.

In a service oriented architecture, each hardware or software component is exposed to the others as a “service”. A service exposes its functionality through a well defined interface that abstracts all the implementation details and that is usually available through the network.

The most commonly known example of a service oriented architecture are the Web Services technologies that are based on web standards such as HTTP and XML. Semantic Web Services proposes to use knowledge representation methods such as ontologies to give some semantic to services functionalities. Semantic description of services makes it possible to improve the interoperability between services designed by different persons or vendors.

Taken out of the box, most SOA implementations have some “defects” preventing their adoption. Web services, due to their name, are perceived as being only for the “web” and also as having a notable performance overhead. Other implementations such as various propositions around the Java virtual machine, often requires to use a particular programming language or are not distributed. Intelligent environments involves many specialist and a hard constraint on the programming language can be a real barrier to SOA adoption.

The PRIMA project has developed OMiSCID, a middleware for service oriented architectures that addresses the particular problematics of intelligent environments. OMiSCID has emerged as an effective tool for unifying access to functionalities provided from the lowest abstraction level components (camera image acquisition, image processing) to abstract services such as activity modeling and personal assistant. OMiSCID has facilitated cooperation by experts from within the PRIMA project as well as in projects with external partners.

3.3. Robust view-invariant Computer Vision

Local Appearance, Affine Invariance, Receptive Fields

3.3.1. Summary

A long-term grand challenge in computer vision has been to develop a descriptor for image information that can be reliably used for a wide variety of computer vision tasks. Such a descriptor must capture the information in an image in a manner that is robust to changes the relative position of the camera as well as the position, pattern and spectrum of illumination.

Members of PRIMA have a long history of innovation in this area, with important results in the area of multi-resolution pyramids, scale invariant image description, appearance based object recognition and receptive field histograms published over the last 20 years. The group has most recently developed a new approach that extends scale invariant feature points for the description of elongated objects using scale invariant ridges. PRIMA has worked with ST Microelectronics to embed its multi-resolution receptive field algorithms into low-cost mobile imaging devices for video communications and mobile computing applications.

3.3.2. Detailed Description

The visual appearance of a neighbourhood can be described by a local Taylor series [48]. The coefficients of this series constitute a feature vector that compactly represents the neighbourhood appearance for indexing and matching. The set of possible local image neighbourhoods that project to the same feature vector are referred to as the "Local Jet". A key problem in computing the local jet is determining the scale at which to evaluate the image derivatives.

Lindeberg [50] has described scale invariant features based on profiles of Gaussian derivatives across scales. In particular, the profile of the Laplacian, evaluated over a range of scales at an image point, provides a local description that is "equi-variant" to changes in scale. Equi-variance means that the feature vector translates exactly with scale and can thus be used to track, index, match and recognize structures in the presence of changes in scale.

A receptive field is a local function defined over a region of an image [56]. We employ a set of receptive fields based on derivatives of the Gaussian functions as a basis for describing the local appearance. These functions resemble the receptive fields observed in the visual cortex of mammals. These receptive fields are applied to color images in which we have separated the chrominance and luminance components. Such functions are easily normalized to an intrinsic scale using the maximum of the Laplacian [50], and normalized in orientation using direction of the first derivatives [56].

The local maxima in x and y and scale of the product of a Laplacian operator with the image at a fixed position provides a "Natural interest point" [52]. Such natural interest points are salient points that may be robustly detected and used for matching. A problem with this approach is that the computational cost of determining intrinsic scale at each image position can potentially make real-time implementation unfeasible.

A vector of scale and orientation normalized Gaussian derivatives provides a characteristic vector for matching and indexing. The oriented Gaussian derivatives can easily be synthesized using the "steerability property" [39] of Gaussian derivatives. The problem is to determine the appropriate orientation. In earlier work by PRIMA members Colin de Verdiere [31], Schiele [56] and Hall [43], proposed normalising the local jet independently at each pixel to the direction of the first derivatives calculated at the intrinsic scale. This results for many view invariant image recognition tasks are described in the next section.

Key results in this area include

- Fast, video rate, calculation of scale and orientation for image description with normalized chromatic receptive fields [34].
- Robust visual features for face tracking [41], [40].
- Direct computation of time to collision over the entire visual field using rate of change of intrinsic scale [54].

We have achieved video rate calculation of scale and orientation normalized Gaussian receptive fields using an O(N) pyramid algorithm [34]. This algorithm has been used to propose an embedded system that provides real time detection and recognition of faces and objects in mobile computing devices.

Applications have been demonstrated for detection, tracking and recognition of faces as well detection of emotions and posture at video rates.

3.4. Perception for Social Interaction

Affective Computing, Perception for social interaction.

Current research on perception for interaction primarily focuses on recognition and communication of linguistic signals. However, most human-to-human interaction is non-verbal and highly dependent on social context. A technology for natural interaction requires abilities to perceive and assimilate non-verbal social signals, to understand and predict social situations, and to acquire and develop social interaction skills.

The overall goal of this research program is to provide the scientific and technological foundations for systems that observe and interact with people in a polite, socially appropriate manner. We address these objectives with research activities in three interrelated areas:

- Multimodal perception for social interactions.
- Learning models for context aware social interaction, and
- Context aware systems and services.

Our approach to each of these areas is to draw on models and theories from the cognitive and social sciences, human factors, and software architectures to develop new theories and models for computer vision and multimodal interaction. Results will be developed, demonstrated and evaluated through the construction of systems and services for polite, socially aware interaction in the context of smart habitats.

3.4.1. Detailed Description

First part of our work on perception for social interaction has concentrated on measuring the physiological parameters of Valence, Arousal and Dominance using visual observation form environmental sensors as well as observation of facial expressions.

People express and feel emotions with their face. Because the face is both externally visible and the seat of emotional expression, facial expression of emotion plays a central role in social interaction between humans. Thus visual recognition of emotions from facial expressions is a core enabling technology for any effort to adapt systems for social interaction.

Constructing a technology for automatic visual recognition of emotions requires solutions to a number of hard challenges. Emotions are expressed by coordinated temporal activations of 21 different facial muscles assisted by a number of additional muscles. Activations of these muscles are visible through subtle deformations in the surface structure of the face. Unfortunately, this facial structure can be masked by facial markings, makeup, facial hair, glasses and other obstructions. The exact facial geometry, as well as the coordinated expression of muscles is unique to each individual. In additions, these deformations must be observed and measured under a large variety of illumination conditions as well as a variety of observation angles. Thus the visual recognition of emotions from facial expression remains a challenging open problem in computer vision.

Despite the difficulty of this challenge, important progress has been made in the area of automatic recognition of emotions from face expressions. The systematic cataloging of facial muscle groups as facial action units by Ekman [38] has let a number of research groups to develop libraries of techniques for recognizing the elements of the FACS coding system [30]. Unfortunately, experiments with that system have revealed that the system is very sensitive to both illumination and viewing conditions, as well as the difficulty in interpreting the resulting activation levels as emotions. In particular, this approach requires a high-resolution image with a high signal-to-noise ratio obtained under strong ambient illumination. Such restrictions are not compatible with the mobile imaging system used on tablet computers and mobile phones that are the target of this effort.

As an alternative to detecting activation of facial action units by tracking individual face muscles, we propose to measure physiological parameters that underlie emotions with a global approach. Most human emotions can be expressed as trajectories in a three dimensional space whose features are the physiological parameters of Pleasure-Displeasure, Arousal-Passivity and Dominance-Submission. These three physiological parameters can be measured in a variety of manners including on-body accelerometers, prosody, heart-rate, head movement and global face expression.

In our work, we address the recognition of social behaviours multimodal information. These are unconscious inmate cognitive processes that are vital to human communication and interaction. Recognition of social behaviours enables anticipation and improves the quality of interaction between humans. Among social behaviours, we have focused on engagement, the expression of intention for interaction. During the engagement phase, many non-verbal signals are used to communicate the intention to engage to the partner [58]. These include posture, gaze, spatial information, gestures, and vocal cues.

For example, within the context of frail or elderly people at home, a companion robot must also be able to detect the engagement of humans in order to adapt their responses during interaction with humans to increase their acceptability. Classical approaches for engagement with robots use spatial information such as human position and speed, human-robot distance and the angle of arrival. Our believe is that uni-modal methods may be suitable for static display [59] and robots in wide space area [49] but not for home environments. In an apartment, relative spatial information of people and robot are not as discriminative as in an open space. Passing by the robot in a corridor should not lead to an engagement detection, and possible socially inappropriate behaviour by the robot.

In our experiments, we used a kompai robot from Robosoft [29]. As an alternative to wearable physiological sensors (such as pulse bracelet Cardiocam, etc.) we integrate multimodal features using a Kinect sensor (see figure 1). In addition of the spatial cues from the laser telemeter, one can use new multimodal features based on persons and skeletons tracking, sound localization, etc. Some of these new features are inspired from results in cognitive science domain [55].

Figure 1. On the left image, one can see the telemeter range in red, the foot detection (blue spot) and the angle view from the Kinect (in green). the middle and right image represent RGB camera in depth view from the Kinect.

Our multimodal approach has been confronted to a robot centered dataset for multimodal social signal processing recorded in a home-like environment [36]. The evaluation on our corpus highlights its robustness and validates use of such technique in real environment. Experimental validation shows that the use of multimodal sensors gives better results than only spatial features (50% of error reduction). Our experimentations also confirm results from [55]: relative shoulder rotation, speed and facing visage are among crucial features for engagement detection.

3.5. End User control of Smart Environments

End users programming, smart home, smart environment

Pervasive computing promises unprecedented empowerment from the flexible and robust combination of software services with the physical world. Software researchers assimilate this promise as system autonomy where users are conveniently kept out of the loop. Their hypothesis is that services, such as music playback and calendars, are developed by service providers and pre-assembled by software designers to form new service frontends. Their scientific challenge is then to develop secure, multiscale, multi-layered, virtualized infrastructures that guarantee service front-end continuity. Although service continuity is desirable in many circumstances, end users, with this interpretation of ubiquitous computing, are doomed to behave as mere consumers, just like with conventional desktop computing.

Another interpretation of the promises of ubiquitous computing, is the empowerment of end users with tools that allow them to create and reshape their own interactive spaces. Our hypothesis is that end users are willing to shape their own interactive spaces by coupling smart artifacts, building imaginative new functionality that were not anticipated by system designers. A number of tools and techniques have been developed to support this view such as CAMP [57] or iCAP [37].

We are investigating an End-User Programming (EUP) approach to give the control back to the inhabitants. In our vision, smart homes will be incrementally equiped with sensors, actuators and services by inhabitants themselves. Our research programm therefore focus on tools and languages to enable inhabitants in activities related to EUP for Smart Homes :

- Installation and maintenance of devices and services. This may imply having facilities to attribute names.
- Visualizing and controling of the Smart Habitat.
- Programming and testing. This imply one or more programming languages and programming environment which could rely on the previous point. The programming language is especially important. Indeed, in the context of the Smart Homes, End-User Programms are most likely to be routines in the sens of [35] than procedure in the sense of traditionnal programming languages.
- Detecting and solving conflicts related to contradictory programs or goals.

REVES Project-Team

3. Research Program

3.1. Plausible Rendering

We consider plausible rendering to be a first promising research direction, both for images and for sound. Recent developments, such as point rendering, image-based modeling and rendering, and work on the simulation of aging indicate high potential for the development of techniques which render *plausible* rather than extremely accurate images. In particular, such approaches can result in more efficient renderings of very complex scenes (such as outdoors environments). This is true both for visual (image) and sound rendering. In the case of images, such techniques are naturally related to image- or point-based methods. It is important to note that these models are becoming more and more important in the context of network or heterogeneous rendering, where the traditional polygon-based approach is rapidly reaching its limits. Another research direction of interest is realistic rendering using simulation methods, both for images and sound. In some cases, research in these domains has reached a certain level of maturity, for example in the case of lighting and global illumination. For some of these domains, we investigate the possibility of technology transfer with appropriate partners. Nonetheless, certain aspects of these research domains, such as visibility or high-quality sound still have numerous and interesting remaining research challenges.

3.1.1. Alternative representations for complex geometry

The key elements required to obtain visually rich simulations, are sufficient geometric detail, textures and lighting effects. A variety of algorithms exist to achieve these goals, for example displacement mapping, that is the displacement of a surface by a function or a series of functions, which are often generated stochastically. With such methods, it is possible to generate convincing representations of terrains or mountains, or of non-smooth objects such as rocks. Traditional approaches used to represent such objects require a very large number of polygons, resulting in slow rendering rates. Much more efficient rendering can be achieved by using point or image based rendering, where the number of elements used for display is view- or image resolution-dependent, resulting in a significant decrease in geometric complexity. Such approaches have very high potential. For example, if all object can be rendered by points, it could be possible to achieve much higher quality local illumination or shading, using more sophisticated and expensive algorithms, since geometric complexity will be reduced. Such novel techniques could lead to a complete replacement of polygon-based rendering for complex scenes. A number of significant technical challenges remain to achieve such a goal, including sampling techniques which adapt well to shading and shadowing algorithms, the development of algorithms and data structures which are both fast and compact, and which can allow interactive or real-time rendering. The type of rendering platforms used, varying from the high-performance graphics workstation all the way to the PDA or mobile phone, is an additional consideration in the development of these structures and algorithms. Such approaches are clearly a suitable choice for network rendering, for games or the modelling of certain natural object or phenomena (such as vegetation, e.g. Figure 1 , or clouds). Other representations merit further research, such as image or video based rendering algorithms, or structures/algorithms such as the "render cache" [31], which we have developed in the past, or even volumetric methods. We will take into account considerations related to heterogeneous rendering platforms, network rendering, and the appropriate choices depending on bandwith or application. Point- or image-based representations can also lead to novel solutions for capturing and representing real objects. By combining real images, sampling techniques and borrowing techniques from other domains (e.g., computer vision, volumetric imaging, tomography etc.) we hope to develop representations of complex natural objects which will allow rapid rendering. Such approaches are closely related to texture synthesis and image-based modeling. We believe that such methods will not replace 3D (laser or range-finger) scans, but could be complementary, and represent a simpler and lower cost alternative for certain applications (architecture, archeology etc.). We are also investigating methods for adding "natural appearance" to synthetic objects. Such approaches include *weathering* or *aging* techniques,

based on physical simulations [21], but also simpler methods such as accessibility maps [28]. The approaches we intend to investigate will attempt to both combine and simplify existing techniques, or develop novel approaches founded on generative models based on observation of the real world.

3.1.2. Plausible audio rendering

Similar to image rendering, plausible approaches can be designed for audio rendering. For instance, the complexity of rendering high order reflections of sound waves makes current geometrical approaches inappropriate. However, such high order reflections drive our auditory perception of "reverberation" in a virtual environment and are thus a key aspect of a plausible audio rendering approach. In complex environments, such as cities, with a high geometrical complexity, hundreds or thousands of pedestrians and vehicles, the acoustic field is extremely rich. Here again, current geometrical approaches cannot be used due to the overwhelming number of sound sources to process. We study approaches for statistical modeling of sound scenes to efficiently deal with such complex environments. We also study perceptual approaches to audio rendering which can result in high efficiency rendering algorithms while preserving visual-auditory consistency if required.

Figure 1. Plausible rendering of an outdoors scene containing points, lines and polygons [20], representing a scene with trees, grass and flowers. We can achieve 7-8 frames per second compared to tens of seconds per image using standard polygonal rendering.

3.2. High Quality Rendering Using Simulation

3.2.1. Non-diffuse lighting

A large body of global illumination research has concentrated on finite element methods for the simulation of the diffuse component and stochastic methods for the non-diffuse component. Mesh-based finite element approaches have a number of limitations, in terms of finding appropriate meshing strategies and form-factor calculations. Error analysis methodologies for finite element and stochastic methods have been very different in the past, and a unified approach would clearly be interesting. Efficient rendering, which is a major advantage of finite element approaches, remains an overall goal for all general global illumination research. For certain cases, stochastic methods can be efficient for all types of light transfers, in particular if we require a view-dependent solution. We are also interested both in *pure* stochastic methods, which do not use finite element techniques. Interesting future directions include filtering for improvement of final image quality as well as beam tracing type approaches [29] which have been recently developed for sound research.

3.2.2. Visibility and Shadows

Visibility calculations are central to all global illumination simulations, as well as for all rendering algorithms of images and sound. We have investigated various global visibility structures, and developed robust solutions for scenes typically used in computer graphics. Such analytical data structures [25], [24], [23] typically have robustness or memory consumption problems which make them difficult to apply to scenes of realistic size. Our solutions to date are based on general and flexible formalisms which describe all visibility event in terms of generators (vertices and edges); this approach has been published in the past [22]. Lazy evaluation, as well as hierarchical solutions, are clearly interesting avenues of research, although are probably quite application dependent.

3.2.3. Radiosity

For purely diffuse scenes, the radiosity algorithm remains one of the most well-adapted solutions. This area has reached a certain level of maturity, and many of the remaining problems are more technology-transfer oriented We are interested in interactive or real-time renderings of global illumination simulations for very complex scenes, the "cleanup" of input data, the use of application-dependent semantic information and mixed representations and their management. Hierarchical radiosity can also be applied to sound, and the ideas used in clustering methods for lighting can be applied to sound.

3.2.4. High-quality audio rendering

Our research on high quality audio rendering is focused on developing efficient algorithms for simulations of geometrical acoustics. It is necessary to develop techniques that can deal with complex scenes, introducing efficient algorithms and data structures (for instance, beam-trees [26] [29]), especially to model early reflections or diffractions from the objects in the environment. Validation of the algorithms is also a key aspect that is necessary in order to determine important acoustical phenomena, mandatory in order to obtain a high-quality result. Recent work by Nicolas Tsingos at Bell Labs [27] has shown that geometrical approaches can lead to high quality modeling of sound reflection and diffraction in a virtual environment (Figure 2). We will pursue this research further, for instance by dealing with more complex geometry (e.g., concert hall, entire building floors).

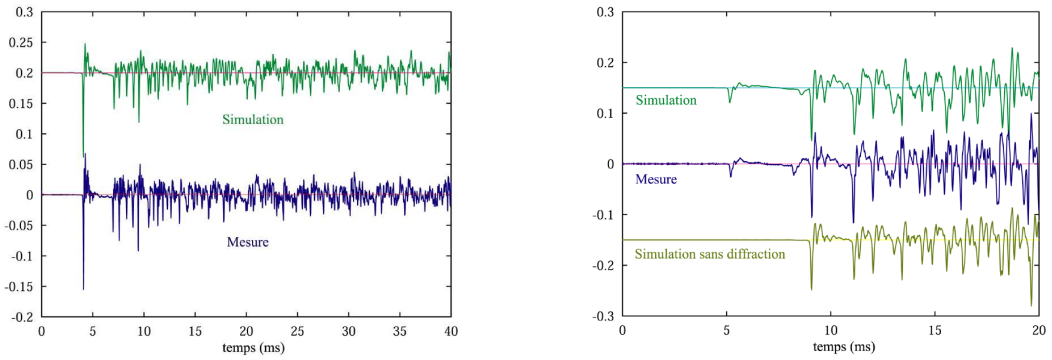


Figure 2. A comparison between a measurement (left) of the sound pressure in a given location of the "Bell Labs Box", a simple test environment built at Bell Laboratories, and a high-quality simulation based on a beam-tracing engine (right). Simulations include effects of reflections off the walls and diffraction off a panel introduced in the room.

Finally, several signal processing issues remain in order to properly and efficiently restitute a 3D soundfield to the ears of the listener over a variety of systems (headphones, speakers). We would like to develop an open

and general-purpose API for audio rendering applications. We already completed a preliminary version of a software library: AURELI [30].

RITS Team

3. Research Program

3.1. Vehicle guidance and autonomous navigation

Participants: Zayed Alsayed, Guillaume Bresson, David Gonzalez Bautista, Wei-Lin Ku, Mohamed Marouf, Pierre Merdrignac, Vicente Milanes Montero, Fawzi Nashashibi, Joshué Pérez Rastelli, Plamen Petrov, Evangeline Pollard, Oyunchimeg Shagdar, Guillaume Trehard, Anne Verroust-Blondet.

There are three basic ways to improve the safety of road vehicles and these ways are all of interest to the project-team. The first way is to assist the driver by giving him better information and warning. The second way is to take over the control of the vehicle in case of mistakes such as inattention or wrong command. The third way is to completely remove the driver from the control loop.

All three approaches rely on information processing. Only the last two involve the control of the vehicle with actions on the actuators, which are the engine power, the brakes and the steering. The research proposed by the project-team is focused on the following elements:

- perception of the environment,
- planning of the actions,
- real-time control.

3.1.1. Perception of the road environment

Participants: Zayed Alsayed, Guillaume Bresson, Wei-Lin Ku, Pierre Merdrignac, Fawzi Nashashibi, Joshué Pérez Rastelli, Evangeline Pollard, Guillaume Trehard, Anne Verroust-Blondet.

Either for driver assistance or for fully automated guided vehicle purposes, the first step of any robotic system is to perceive the environment in order to assess the situation around itself. Proprioceptive sensors (accelerometer, gyrometer,...) provide information about the vehicle by itself such as its velocity or lateral acceleration. On the other hand, exteroceptive sensors, such as video camera, laser or GPS devices, provide information about the environment surrounding the vehicle or its localization. Obviously, fusion of data with various other sensors is also a focus of the research.
The following topics are already validated or under development in our team:

- relative ego-localization with respect to the infrastructure, i.e. lateral positioning on the road can be obtained by mean of vision (lane markings) and the fusion with other devices (e.g. GPS);
- global ego-localization by considering GPS measurement and proprioceptive information, even in case of GPS outage;
- road detection by using lane marking detection and navigable free space;
- detection and localization of the surrounding obstacles (vehicles, pedestrians, animals, objects on roads, etc.) and determination of their behavior can be obtained by the fusion of vision, laser or radar based data processing;
- simultaneous localization and mapping as well as mobile object tracking using laser-based and stereovision-based (SLAMMOT) algorithms.

This year was the opportunity to focus on two particular topics: SLAMMOT-based techniques for grid-based environment modeling using laser sensors, and belief-based SLAM techniques for vehicle navigation.

3.1.2. 3D environment representation

Participants: Zayed Alsayed, Guillaume Bresson, Fawzi Nashashibi.

In the past few years, we have been focusing on the Disparity map estimation as a mean to obtain dense 3D mapping of the environment. Moreover, many autonomous vehicle navigation systems have adopted stereo vision techniques to construct disparity maps as a basic obstacle detection and avoidance mechanism. Two different approaches where investigated: the Fly algorithm, and the stereo vision for 3D representation.

In the first approach, the Fly algorithm is an evolutionary optimization applied to stereovision and mobile robotics. Its advantage relies on its precision and its acceptable costs (computation time and resources). In the second approach, originality relies on computing the disparity field by directly formulating the problem as a constrained optimization problem in which a convex objective function is minimized under convex constraints. These constraints arise from prior knowledge and the observed data. The minimization process is carried out over the feasibility set and with a suitable regularization constraint: the Total Variation information, which avoids oscillations while preserving field discontinuities around object edges. Although successfully applied to real-time pedestrian detection using a vehicle mounted stereohead (see LOVe project), this technique could not be used for other robotics applications such as scene modeling, visual SLAM, etc. The need is for a dense 3D representation of the environment obtained with an appropriate precision and acceptable costs (computation time and resources).

Stereo vision is a reliable technique for obtaining a 3D scene representation through a pair of left and right images and it is effective for various tasks in road environments. The most important problem in stereo image processing is to find corresponding pixels from both images, leading to the so-called disparity estimation. Many autonomous vehicle navigation systems have adopted stereo vision techniques to construct disparity maps as a basic obstacle detection and avoidance mechanism. We are presently working on an original stereo-vision based SLAM technique, which aimed at reconstructing current surroundings through on-the-fly real-time localization of tens of thousands of interest points. This development should also allow detection and tracking of moving objects [0], and is built on linear algebra (through Inria's Eigen library), using the RANSAC algorithm and multi-target tracking techniques, to quote a few.

This technique complements another laser based SLAMMOT technique developed since few years and extensively validated in large scale demonstrations for indoor and outdoor robotics applications. This technique has proved its efficiency in terms of cost, accuracy and reliability.

3.1.3. Cooperative Multi-sensor data fusion

Participants: Pierre Merdrignac, Fawzi Nashashibi, Evangeline Pollard, Oyunchimeg Shagdar.

Since data are noisy, inaccurate and can also be unreliable or unsynchronized, the use of data fusion techniques is required in order to provide the most accurate situation assessment as possible to perform the perception task. RITS team worked a lot on this problem in the past, but is now focusing on collaborative perception approach. Indeed, the use of vehicle-to-vehicle or vehicle-to-infrastructure communications allows an improved on-board reasoning since the decision is made based on an extended perception.

As a direct consequence of the electronics broadly used for vehicular applications, communication technologies are now being adopted as well. In order to limit injuries and to share safety information, research in driving assistance system is now orientating toward the cooperative domain. Advanced Driver Assistance System (ADAS) and Cybercars applications are moving towards vehicle-infrastructure cooperation. In such scenario, information from vehicle based sensors, roadside based sensors and a priori knowledge is generally combined thanks to wireless communications to build a probabilistic spatio-temporal model of the environment. Depending on the accuracy of such model, very useful applications from driver warning to fully autonomous driving can be performed.

The Collaborative Perception Framework (CPF) is a combined hardware/software approach that permits to see remote information as its own information. Using this approach, a communicant entity can see another remote entity software objects as if it was local, and a sensor object, can see sensor data of others entities as its own sensor data. Last year we developed the basic hardware modules that ensure the well functioning of the embedded architecture including perception sensors, communication devices and processing tools.

[0] http://www.youtube.com/watch?v=obH9Z2uOMBI

Finally, since vehicle localization (ground vehicles) is an important task for intelligent vehicle systems, vehicle cooperation may bring benefits for this task. A new cooperative multi-vehicle localization method using split covariance intersection filter was developed during the year 2012, as well as a cooperative GPS data sharing method.

In the first method, each vehicle estimates its own position using a SLAM approach. In parallel, it estimates a decomposed group state, which is shared with neighboring vehicles; the estimate of the decomposed group state is updated with both the sensor data of the ego-vehicle and the estimates sent from other vehicles; the covariance intersection filter which yields consistent estimates even facing unknown degree of inter-estimate correlation has been used for data fusion.

In the second GPS data sharing method, a new collaborative localization method is proposed. On the assumption that the distance between two communicative vehicles can be calculated with a good precision, cooperative vehicle are considered as additional satellites into the user position calculation by using iterative methods. In order to limit divergence, some filtering process is proposed: Interacting Multiple Model (IMM) is used to guarantee a greater robustness in the user position estimation.

Accidents between vehicles and pedestrians (including cyclists) often result in fatality or at least serious injury for pedestrians, showing the need of technology to protect vulnerable road users. Vehicles are now equipped with many sensors in order to model their environment, to localize themselves, detect and classify obstacles, etc. They are also equipped with communication devices in order to share the information with other road users and the environment. The goal of this work is to develop a cooperative perception and communication system, which merges information coming from the communications device and obstacle detection module to improve the pedestrian detection, tracking, and hazard alarming.

Pedestrian detection is performed by using a perception architecture made of two sensors: a laser scanner and a CCD camera. The laser scanner provides a first hypothesis on the presence of a pedestrian-like obstacle while the camera performs the real classification of the obstacle in order to identify the pedestrian(s). This is a learning-based technique exploiting adaptive boosting (AdaBoost). Several classifiers were tested and learned in order to determine the best compromise between the nature and the number of classifiers and the accuracy of the classification.

3.1.4. Planning and executing vehicle actions

Participants: David Gonzalez Bautista, Mohamed Marouf, Vicente Milanes Montero, Fawzi Nashashibi, Joshué Pérez Rastelli, Plamen Petrov.

From the understanding of the environment, thanks to augmented perception, we have either to warn the driver to help him in the control of his vehicle, or to take control in case of a driverless vehicle. In simple situations, the planning might also be quite simple, but in the most complex situations we want to explore, the planning must involve complex algorithms dealing with the trajectories of the vehicle and its surroundings (which might involve other vehicles and/or fixed or moving obstacles). In the case of fully automated vehicles, the perception will involve some map building of the environment and obstacles, and the planning will involve partial planning with periodical recomputation to reach the long term goal. In this case, with vehicle to vehicle communications, what we want to explore is the possibility to establish a negotiation protocol in order to coordinate nearby vehicles (what humans usually do by using driving rules, common sense and/or non verbal communication). Until now, we have been focusing on the generation of geometric trajectories as a result of a maneuver selection process using grid-based rating technique or fuzzy technique. For high speed vehicles, Partial Motion Planning techniques we tested, revealed their limitations because of the computational cost. The use of quintic polynomials we designed, allowed us to elaborate trajectories with different dynamics adapted to the driver profile. These trajectories have been implemented and validated in the JointSystem demonstrator of the German Aerospace Center (DLR) used in the European project HAVEit, as well as in RITS's electrical vehicle prototype used in the French project ABV. HAVEit was also the opportunity for RITS to take in charge the implementation of the Co-Pilot system which processes perception data in order to elaborate the high level command for the actuators. These trajectories were also validated on RITS's cybercars. However, for the low speed cybercars that have pre-defined itineraries and basic maneuvers, it was necessary

to develop a more adapted planning and control system. Therefore, we have developed a nonlinear adaptive control for automated overtaking maneuver using quadratic polynomials and Lyapunov function candidate and taking into account the vehicles kinematics. For the global mobility systems we are developing, the control of the vehicles includes also advanced platooning, automated parking, automated docking, etc. For each functionality a dedicated control algorithm was designed (see publication of previous years). Today, RITS is also investigating the opportunity of fuzzy-based control for specific maneuvers. First results have been recently obtained for reference trajectories following in roundabouts and normal straight roads.

3.2. V2V and V2I Communications for ITS

Participants: Thierry Ernst, Oyunchimeg Shagdar, Gérard Le Lann, Younes Bouchaala, Pierre Merdrignac, Ines Ben Jemaa, Mohammad Abu Alhoul, Fawzi Nashashibi, Arnaud de La Fortelle.

Wireless communications are expected to play an important role for road safety, road efficiency, and comfort of road users. Road safety applications often require highly responsive and reliable information exchange between neighboring vehicles in any road density condition. Because the performance of the existing radio communications technology largely degrades with the increase of the node density, the challenge of designing wireless communications for safety applications is enabling reliable communications in highly dense scenarios. Targeting this issue, RITS has been working on medium access control design and visible light communications, especially for highly dense scenarios. The works have been carried out considering the vehicle behavior such as vehicle merging and vehicle platooning.

Unlike many of the road safety applications, the applications regarding road efficiency and comfort of road users, on the other hand, often require connectivity to the Internet. Based on our expertise in both Internet-based communications in the mobility context and in ITS, we are now investigating the use of IPv6 (Internet Protocol version 6 which is going to replace the current version, IPv4, in a few years from now) for vehicular communications, in a combined architecture allowing both V2V and V2I.

The wireless channel and the topology dynamics need to be studied when understanding the dynamics and designing efficient communications mechanisms. Targeting this issue, we have been working on channel modeling for both radio and visible light communications, and design of communications mechanisms especially for security, service discovery, multicast and geocast message delivery, and access point selection.

Below follows a more detailed description of the related research issues.

3.2.1. Geographic multicast addressing and routing

Participants: Ines Ben Jemaa, Oyunchimeg Shagdar, Thierry Ernst, Arnaud de La Fortelle.

Many ITS applications such as fleet management require multicast data delivery. Existing work on this subject tackles mainly the problems of IP multicasting inside the Internet or geocasting in the VANETs. To enable Internet-based multicast services for VANETs, we introduced a framework that:
i) defines a distributed and efficient geographic multicast auto-addressing mechanism to ensure vehicular multicast group reachability through the infrastructure network,
ii) introduces a simplified approach that locally manages the group membership and distributes the packets among themto allow simple and efficient data delivery.

3.2.2. Platooning control using visible light communications

Participants: Mohammad Abu Alhoul, Mohamed Marouf, Oyunchimeg Shagdar, Fawzi Nashashibi.

The main purpose of our research is to propose and test new successful supportive communication technology, which can provide stable and reliable communication between vehicles, especially for the platooning scenario. Although VLC technology has a short history in comparison with other communication technologies, the infrastructure availability and the presence of the congestion in wireless communication channels lead to propose VLC technology as a reliable and supportive technology which can takeoff some loads of the wireless radio communication. The first objective of this work is to develop an analytical model of VLC to understand its characteristics and limitations. The second objective is to design vehicle platooning control using VLC. In platooning control, a cooperation between control and communication is strongly required in order to guarantee the platoon's stability (e.g. string stability problem). For this purpose we work on VLC model platooning scenario, to permit for each vehicle the trajectory tracking of the vehicle ahead, altogether with a prescribed inter-vehicle distance and considering all the VLC channel model limitations. The integrated channel model of the main Simulink platooning model will be responsible for deciding the availability of the Line-of-Sight for different trajectory's curvatures, which means the capability of using light communication between each couple of vehicles in the platooning queue. At the same time the model will compute all the required parameters acquired from each vehicle controller.

3.2.3. V2X radio communications for road safety applications

Participants: Mohammad Abu Alhoul, Pierre Merdrignac, Oyunchimeg Shagdar, Fawzi Nashashibi.

While 5.9 GHz radio frequency band is dedicated to ITS applications, the channel and network behaviors in mobile scenarios are not very well known. In this work we theoretically and experimentally study the radio channel characteristics in vehicular networks, especially the radio quality and bandwidth availability. Based on our study, we develop mechanisms for efficient and reliable V2X communications, channel allocation, congestion control, and access point selection, which are especially dedicated to road safety and autonomous driving applications.

3.3. Automated driving, intelligent vehicular networks, and safety

Participant: Gérard Le Lann.

Intelligent vehicular networks (IVNs) are one constituent of ITS. IVNs encompass "clusters", platoons and vehicular ad-hoc networks comprising automated and cooperative vehicles. A basic principle that underlies our work is minimal reliance on road-side infrastructures for solving those open problems arising with IVNs. For example, V2V communications only are considered. Trivially, if one can solve a problem P considering V2V communications only, then P is solved with the help of V2I communications, whereas the converse is not true. Moreover, safety in the course of risk-prone maneuvers is our central concern. Since safety-critical (SC) scenarios may develop anytime anywhere, it is impossible to assume that there is always a road-side unit in the vicinity of those vehicles involved in a hazardous situation.

3.3.1. Cohorts and groups – Novel constructs for safe IVNs

The automated driving function rests on two radically different sets of solutions, one set encompassing signal processing and robotics (SPR), the other one encompassing vehicular communications and networking (VCN). In addition to being used for backing a failing SPR solution, VCN solutions have been originally proposed for "augmenting" the capabilities offered by SPR solutions, which are line-of-sight technologies, i.e. limited by obstacles. Since V2V omni-directional radio communications that are being standardized (IEEE 802.11p / WAVE) have ranges in the order of 250 m, it is interesting to prefix risk-prone maneuvers with the exchange of SC-messages. Roles being assigned prior to initiating physical maneuvers, the SPR solutions are invoked under favorable conditions, safer than when vehicles have not agreed on "what to do" ahead of time.

VCN solutions shall belong to two categories: V2V omni-directional (360°) communications and unidirectional communications, implemented out of very-short range antennas of very small beam-width. This has led to the concept of neighbor-to-neighbor (N2N) communications, whereby vehicles following each other on a given lane can exchange periodic beacons and event-driven messages.

Vehicle motions on roads and highways obey two different regimes. First, stationary regimes, where inter-vehicular spacing, acceleration and deceleration rates (among other parameters), match specified bounds. This, combined with N2N communications, has led to the concept of cohorts, where safety is not at stake provided that no violation of bounds occurs. Second, transitory regimes, where some of these bounds are violated (e.g., sudden braking – the "brick wall" paradigm), or where vehicles undertake risk-prone maneuvers such as lane changes, resulting into SC scenarios. Reasoning about SC scenarios has led to the concept of groups. Cohorts and groups have been introduced in [6].

3.3.2. Cohorts, N2N communications, and safety in the presence of telemetry failures

In [6] we show how periodic N2N beaconing serves to withstand failures of directional telemetry devices. Worst-case bounds on safe inter-vehicular spacing are established analytically (simulations cannot be used for establishing worst-case bounds). A result of practical interest is the ability to answer the following question: "vehicles move at high speed in a cohort formation; if in a platoon formation, spacing would be in the order of 3 m; what is the additional safe spacing in a cohort?" With a N2N beaconing period in the range of 100-200 ms, the additional spacing is much less than 1 m. Failure of a N2N communication link translates into a cohort split, one of the vehicles impaired becoming the tail of a cohort, and its (impaired) follower becoming the head of a newly formed cohort. The number of vehicles in a cohort has an upper bound, and the inter-cohort spacing has a lower bound.

3.3.3. Groups, cohorts, and fast reliable V2V Xcasting in the presence of message losses

Demonstrating safety involves establishing strict timeliness ("real-time") properties under worst-case conditions (traffic density, failure rates, radio interference ranges). As regards V2V message passing, this requirement translates into two major problems:

- TBD: time-bounded delivery of V2V messages exchanged among vehicles that undertake SC maneuvers, despite high message loss ratios.
- TBA: time-bounded access to a radio channel in open ad hoc, highly mobile, networks of vehicles, some vehicles undertaking SC maneuvers, despite high contention.

Groups and cohorts have proved to be essential constructs for devising a solution for problem TBD. Vehicles involved in a SC scenario form a group where a 3-way handshake is unfolded so as to reach an agreement regarding roles and adjusted motions. A 3-way handshake consists in 3 rounds of V2V Xcasting of SC messages, round 1 being a Geocast, round 2 being a Convergecast, and round 3 being a Multicast. Worst-case time bound for completing a 3-way handshake successfully is in the order of 200 ms, under worst-case conditions. It is well known that message losses are the dominant cause of failures in mobile wireless networks, which raises the following problem with the Xcasting of SC messages. If acknowledgments are not used, it is impossible to predict probabilities for successful deliveries, which is antagonistic with demonstrating safety. Asking for acknowledgments is a non solution. Firstly, by definition, vehicles that are to be reached by a Geocast are unknown to a sender. How can a sender know which acknowledgments to wait for? Secondly, repeating a SC message that has been lost on a radio channel does not necessarily increases chances of successful delivery. Indeed, radio interferences (causing the first transmission loss) may well last longer than 200 ms (or seconds). To be realistic, one is led to consider a novel and extremely powerful (adversary) failure model (denoted Ω), namely the restricted unbounded omission model, whereby messages meant to circulate on f out of n radio links are "erased" by the adversary (the same f links), ad infinitum. Moreover, we have assumed message loss ratios f/n as high as $2/3$. This is the setting we have considered in [59], where we present a solution for the fast (less than 200 ms) reliable (in the presence of Ω) multipoint communications problem TBD. The solution consists in a suite of Xcast protocols (the Zebra suite) and proxy sets built out of cohorts. Analytical expressions are given for the worst-case time bounds for each of the Zebra protocols.

Surprisingly, while not being originally devised to that end, it turns out that cohorts and groups are essential cornerstones for solving open problem TBA.

3.4. Probabilistic modeling for large transportation systems

Participants: Guy Fayolle, Cyril Furtlehner, Arnaud de La Fortelle, Jean-Marc Lasgouttes.

This activity concerns the modeling of random systems related to ITS, through the identification and development of solutions based on probabilistic methods and more specifically through the exploration of links between large random systems and statistical physics. Traffic modeling is a very fertile area of application for this approach, both for macroscopic (fleet management [4], traffic prediction) and for microscopic (movement of each vehicle, formation of traffic jams) analysis. When the size or volume of structures grows (leading to the so-called "thermodynamic limit"), we study the quantitative and qualitative (performance, speed, stability, phase transitions, complexity, etc.) features of the system.

In the recent years, several directions have been explored.

3.4.1. Traffic reconstruction

Large random systems are a natural part of macroscopic studies of traffic, where several models from statistical physics can be fruitfully employed. One example is fleet management, where one main issue is to find optimal ways of reallocating unused vehicles: it has been shown that Coulombian potentials might be an efficient tool to drive the flow of vehicles. Another case deals with the prediction of traffic conditions, when the data comes from probe vehicles instead of static sensors.

While the widely-used macroscopic traffic flow models are well adapted to highway traffic, where the distance between junction is long (see for example the work done by the NeCS team in Grenoble), our focus is on a more urban situation, where the graphs are much denser. The approach we are advocating here is model-less, and based on statistical inference rather than fundamental diagrams of road segments. Using the Ising model or even a Gaussian Random Markov Field, together with the very popular Belief Propagation (BP) algorithm, we have been able to show how real-time data can be used for traffic prediction and reconstruction (in the space-time domain).

This new use of BP algorithm raises some theoretical questions about the ways the make the belief propagation algorithm more efficient:

- find the best way to inject real-valued data in an Ising model with binary variables [60];
- build macroscopic variables that measure the overall state of the underlying graph, in order to improve the local propagation of information [58];
- make the underlying model as sparse as possible, in order to improve BP convergence and quality [40].

3.4.2. Exclusion processes for road traffic modeling

The focus here is on road traffic modeled as a granular flow, in order to analyze the features that can be explained by its random nature. This approach is complementary to macroscopic models of traffic flow (as done for example in the Opale team at Inria), which rely mainly on ODEs and PDEs to describe the traffic as a fluid.

One particular feature of road traffic that is of interest to us is the spontaneous formation of traffic jams. It is known that systems as simple as the Nagel-Schreckenberg model are able to describe traffic jams as an emergent phenomenon due to interaction between vehicles. However, even this simple model cannot be explicitly analyzed and therefore one has to resort to simulation.

One of the simplest solvable (but non trivial) probabilistic models for road traffic is the exclusion process. It lends itself to a number of extensions allowing to tackle some particular features of traffic flows: variable speed of particles, synchronized move of consecutive particles (platooning), use of geometries more complex than plain 1D (cross roads or even fully connected networks), formation and stability of vehicle clusters (vehicles that are close enough to establish an ad-hoc communication system), two-lane roads with overtaking.

The aspect that we have particularly studied is the possibility to let the speed of vehicle evolve with time. To this end, we consider models equivalent to a series of queues where the pair (service rate, number of customers) forms a random walk in the quarter plane $\mathbb{Z}_+^2$.

Having in mind a global project concerning the analysis of complex systems, we also focus on the interplay between discrete and continuous description: in some cases, this recurrent question can be addressed quite rigorously via probabilistic methods.

We have considered in [57] some classes of models dealing with the dynamics of discrete curves subjected to stochastic deformations. It turns out that the problems of interest can be set in terms of interacting exclusion processes, the ultimate goal being to derive hydrodynamic limits after proper scaling. A seemingly new method is proposed, which relies on the analysis of specific partial differential operators, involving variational calculus and functional integration. Starting from a detailed analysis of the Asymmetric Simple Exclusion Process (ASEP) system on the torus $\mathbb{Z}/n\mathbb{Z}$, the arguments a priori work in higher dimensions (ABC, multi-type exclusion processes, etc), leading to systems of coupled partial differential equations of Burgers' type.

3.4.3. Random walks in the quarter plane $\mathbb{Z}_+^2$

This field remains one of the important *"violon d'Ingres"* in our research activities in stochastic processes, both from theoretical and applied points of view. In particular, it is a building block for models of many communication and transportation systems.

One essential question concerns the computation of stationary measures (when they exist). As for the answer, it has been given by original methods formerly developed in the team (see books and related bibliography). For instance, in the case of small steps (jumps of size one in the interior of $\mathbb{Z}_+^2$), the invariant measure $\{\pi_{i,j}, i,j \geq 0\}$ does satisfy the fundamental functional equation (see [3]):

$$Q(x,y)\pi(x,y) = q(x,y)\pi(x) + \widetilde{q}(x,y)\widetilde{\pi}(y) + \pi_0(x,y). \tag{94}$$

where the unknown generating functions $\pi(x,y), \pi(x), \widetilde{\pi}(y), \pi_0(x,y)$ are sought to be analytic in the region $\{(x,y) \in \mathbb{C}^2 : |x| < 1, |y| < 1\}$, and continuous on their respective boundaries.

The given function $Q(x,y) = \sum_{i,j} p_{i,j} x^i y^j - 1$, where the sum runs over the possible jumps of the walk inside $\mathbb{Z}_+^2$, is often referred to as the *kernel*. Then it has been shown that equation (1) can be solved by reduction to a boundary-value problem of Riemann-Hilbert type. This method has been the source of numerous and fruitful developments. Some recent and ongoing works have been dealing with the following matters.

- *Group of the random walk*. In several studies, it has been noticed that the so-called *group of the walk* governs the behavior of a number of quantities, in particular through its *order*, which is always even. In the case of small jumps, the algebraic curve R defined by $\{Q(x,y) = 0\}$ is either of *genus* 0 (the sphere) or 1 (the torus). In [Fayolle-2011a], when the drift of the random walk is equal to 0 (and then so is the genus), an effective criterion gives the *order* of the group. More generally, it is also proved that whenever the genus is 0, this order is infinite, except precisely for the zero drift case, where finiteness is quite possible. When the *genus* is 1, the situation is more difficult. Recently [55], a criterion has been found in terms of a determinant of order 3 or 4, depending on the arity of the group.
- *Nature of the counting generating functions*. Enumeration of planar lattice walks is a classical topic in combinatorics. For a given set of allowed jumps (or steps), it is a matter of counting the number of paths starting from some point and ending at some arbitrary point in a given time, and possibly restricted to some regions of the plane. A first basic and natural question arises: how many such paths exist? A second question concerns the nature of the associated counting generating functions (CGF): are they rational, algebraic, holonomic (or D-finite, i.e. solution of a linear differential equation with polynomial coefficients)?

  Let $f(i,j,k)$ denote the number of paths in $\mathbb{Z}_+^2$ starting from $(0,0)$ and ending at (i,j) at time k. Then the corresponding CGF

$$F(x,y,z) = \sum_{i,j,k \geq 0} f(i,j,k) x^i y^j z^k \tag{95}$$

satisfies the functional equation

$$K(x,y)F(x,y,z) = c(x)F(x,0,z) + \widetilde{c}(y)F(0,y,z) + c_0(x,y), \tag{96}$$

where z is considered as a time-parameter. Clearly, equations (2) and (1) are of the same nature, and answers to the above questions have been given in [Fayolle-2010].

- *Some exact asymptotics in the counting of walks in* $\mathbb{Z}_+^2$. A new and uniform approach has been proposed about the following problem: *What is the asymptotic behavior, as their length goes to infinity, of the number of walks ending at some given point or domain (for instance one axis)?* The method in [Fayolle-2012] works for *both* finite or infinite groups, and for walks not necessarily restricted to excursions.

3.4.4. Discrete-event simulation for urban mobility

We have developed two simulation tools to study and evaluate the performance of different transportation modes covering an entire urban area.

- one for collective taxis, a public transportation system with a service quality provided will be comparable with that of conventional taxis (system operating with or without reservations, door-to-door services, well adapted itineraries following the current demand, controlling detours and waits, etc.), and with fares set at rates affordable by almost everyone, simply by utilizing previously wasted vehicle capacity;
- the second for a system of self-service cars that can reconfigure themselves into shuttles, therefore creating a multimodal public transportation system; this second simulator is intended to become a generic tool for multimodal transportation.

These two programs use a technique allowing to run simulations in batch mode and analyze the dynamics of the system afterwards.

SEMAGRAMME Project-Team

3. Research Program

3.1. Foundations

The Sémagramme project relies on deep mathematical foundations. We intend to develop models based on well-established mathematics. We seek two main advantages from this approach. On the one hand, by relying on mature theories, we have at our disposal sets of mathematical tools that we can use to study our models. On the other hand, developing various models on a common mathematical background will make them easier to integrate, and will ease the search for unifying principles.

The main mathematical domains on which we rely are formal language theory, symbolic logic, and type theory.

3.1.1. Formal language theory

Formal language theory studies the purely syntactic and combinatorial aspects of languages, seen as sets of strings (or possibly trees or graphs). Formal language theory has been especially fruitful for the development of parsing algorithms for context-free languages. We use it, in a similar way, to develop parsing algorithms for formalisms that go beyond context-freeness. Language theory also appears to be very useful in formally studying the expressive power and the complexity of the models we develop.

3.1.2. Symbolic logic

Symbolic logic (and, more particularly, proof-theory) is concerned with the study of the expressive and deductive power of formal systems. In a rule-based approach to computational linguistics, the use of symbolic logic is ubiquitous. As we previously said, at the level of syntax, several kinds of grammars (generative, categorial...) may be seen as basic deductive systems. At the level of semantics, the meaning of an utterance is capture by computing (intermediate) semantic representations that are expressed as logical forms. Finally, using symbolic logics allows one to formalize notions of inference and entailment that are needed at the level of pragmatics.

3.1.3. Type theory and typed λ-calculus

Among the various possible logics that may be used, Church's simply typed λ-calculus and simple theory of types (a.k.a. higher-order logic) play a central part. On the one hand, Montague semantics is based on the simply typed λ-calculus, and so is our syntax-semantics interface model. On the other hand, as shown by Gallin, [56] the target logic used by Montague for expressing meanings (i.e., his intensional logic) is essentially a variant of higher-order logic featuring three atomic types (the third atomic type standing for the set of possible worlds).

SIROCCO Project-Team

3. Research Program

3.1. Introduction

The research activities on analysis, compression and communication of visual data mostly rely on tools and formalisms from the areas of statistical image modelling, of signal processing, of coding and information theory. However, the objective of better exploiting the Human Visual System (HVS) properties in the above goals also pertains to the areas of perceptual modelling and cognitive science. Some of the proposed research axes are also based on scientific foundations of computer vision (e.g. multi-view modelling and coding). We have limited this section to some tools which are central to the proposed research axes, but the design of complete compression and communication solutions obviously rely on a large number of other results in the areas of motion analysis, transform design, entropy code design, etc which cannot be all described here.

3.2. Parameter estimation and inference

Bayesian estimation, Expectation-Maximization, stochastic modelling

Parameter estimation is at the core of the processing tools studied and developed in the team. Applications range from the prediction of missing data or future data, to extracting some information about the data in order to perform efficient compression. More precisely, the data are assumed to be generated by a given stochastic data model, which is partially known. The set of possible models translates the a priori knowledge we have on the data and the best model has to be selected in this set. When the set of models or equivalently the set of probability laws is indexed by a parameter (scalar or vectorial), the model is said parametric and the model selection resorts to estimating the parameter. Estimation algorithms are therefore widely used at the encoder in order to analyze the data. In order to achieve high compression rates, the parameters are usually not sent and the decoder has to jointly select the model (i.e. estimate the parameters) and extract the information of interest.

3.3. Data Dimensionality Reduction

Manifolds, locally linear embedding, non-negative matrix factorization, principal component analysis

A fundamental problem in many data processing tasks (compression, classification, indexing) is to find a suitable representation of the data. It often aims at reducing the dimensionality of the input data so that tractable processing methods can then be applied. Well-known methods for data dimensionality reduction include principal component analysis (PCA) and independent component analysis (ICA). The methodologies which will be central to several proposed research problems will instead be based on sparse representations, on locally linear embedding (LLE) and on the "non negative matrix factorization" (NMF) framework.

The objective of *sparse representations* is to find a sparse approximation of a given input data. In theory, given $A \in \mathbb{R}^{m \times n}$, $m < n$, and $\mathbf{b} \in \mathbb{R}^m$ with $m << n$ and A is of full rank, one seeks the solution of $\min\{\|\mathbf{x}\|_0 \ : \ A\mathbf{x} = \mathbf{b}\}$, where $\|\mathbf{x}\|_0$ denotes the L_0 norm of x, i.e. the number of non-zero components in z. There exist many solutions x to $Ax = b$. The problem is to find the sparsest, the one for which x has the fewest non zero components. In practice, one actually seeks an approximate and thus even sparser solution which satisfies $\min\{\|\mathbf{x}\|_0 \ : \ \|A\mathbf{x} - \mathbf{b}\|_p \leq \rho\}$, for some $\rho \geq 0$, characterizing an admissible reconstruction error. The norm p is usually 2, but could be 1 or ∞ as well. Except for the exhaustive combinatorial approach, there is no known method to find the exact solution under general conditions on the dictionary A. Searching for this sparsest representation is hence unfeasible and both problems are computationally intractable. Pursuit algorithms have been introduced as heuristic methods which aim at finding approximate solutions to the above problem with tractable complexity.

Non negative matrix factorization (NMF) is a non-negative approximate data representation [0]. NMF aims at finding an approximate factorization of a non-negative input data matrix V into non-negative matrices W and H, where the columns of W can be seen as *basis vectors* and those of H as coefficients of the linear approximation of the input data. Unlike other linear representations like PCA and ICA, the non-negativity constraint makes the representation purely additive. Classical data representation methods like PCA or Vector Quantization (VQ) can be placed in an NMF framework, the differences arising from different constraints being placed on the W and H matrices. In VQ, each column of H is constrained to be unitary with only one non-zero coefficient which is equal to 1. In PCA, the columns of W are constrained to be orthonormal and the rows of H to be orthogonal to each other. These methods of data-dependent dimensionality reduction will be at the core of our visual data analysis and compression activities.

3.4. Perceptual Modelling

Saliency, visual attention, cognition

The human visual system (HVS) is not able to process all visual information of our visual field at once. To cope with this problem, our visual system must filter out irrelevant information and reduce redundant information. This feature of our visual system is driven by a selective sensing and analysis process. For instance, it is well known that the greatest visual acuity is provided by the fovea (center of the retina). Beyond this area, the acuity drops down with the eccentricity. Another example concerns the light that impinges on our retina. Only the visible light spectrum lying between 380 nm (violet) and 760 nm (red) is processed. To conclude on the selective sensing, it is important to mention that our sensitivity depends on a number of factors such as the spatial frequency, the orientation or the depth. These properties are modeled by a sensitivity function such as the Contrast Sensitivity Function (CSF).

Our capacity of analysis is also related to our visual attention. Visual attention which is closely linked to eye movement (note that this attention is called *overt* while the covert attention does not involve eye movement) allows us to focus our biological resources on a particular area. It can be controlled by both top-down (i.e. goal-directed, intention) and bottom-up (stimulus-driven, data-dependent) sources of information [0]. This detection is also influenced by prior knowledge about the environment of the scene [0]. Implicit assumptions related to prior knowledge or beliefs play an important role in our perception (see the example concerning the assumption that light comes from above-left). Our perception results from the combination of prior beliefs with data we gather from the environment. A Bayesian framework is an elegant solution to model these interactions [0]. We define a vector $\overrightarrow{v}_l$ of local measurements (contrast of color, orientation, etc.) and vector $\overrightarrow{v}_c$ of global and contextual features (global features, prior locations, type of the scene, etc.). The salient locations S for a spatial position $\overrightarrow{x}$ are then given by:

$$S(\overrightarrow{x}) = \frac{1}{p(\overrightarrow{v}_l \,|\, \overrightarrow{v}_c)} \times p(s, \overrightarrow{x} \,|\, \overrightarrow{v}_c) \tag{97}$$

The first term represents the bottom-up salience. It is based on a kind of contrast detection, following the assumption that rare image features are more salient than frequent ones. Most of existing computational models of visual attention rely on this term. However, different approaches exist to extract the local visual features as well as the global ones. The second term is the contextual priors. For instance, given a scene, it indicates which parts of the scene are likely the most salient.

[0] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization", Nature 401, 6755, (Oct. 1999), pp. 788-791.
[0] L. Itti and C. Koch, "Computational Modelling of Visual Attention" , Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203, 2001.
[0] J. Henderson, "Regarding scenes", Directions in Psychological Science, vol. 16, pp. 219-222, 2007.
[0] L. Zhang, M. Tong, T. Marks, H. Shan, H. and G.W. Cottrell, "SUN: a Bayesian framework for saliency using natural statistics", Journal of Vision, vol. 8, pp. 1-20, 2008.

3.5. Coding theory

OPTA limit (Optimum Performance Theoretically Attainable), Rate allocation, Rate-Distortion optimization, lossy coding, joint source-channel coding multiple description coding, channel modelization, oversampled frame expansions, error correcting codes.

Source coding and channel coding theory [0] is central to our compression and communication activities, in particular to the design of entropy codes and of error correcting codes. Another field in coding theory which has emerged in the context of sensor networks is Distributed Source Coding (DSC). It refers to the compression of correlated signals captured by different sensors which do not communicate between themselves. All the signals captured are compressed independently and transmitted to a central base station which has the capability to decode them jointly. DSC finds its foundation in the seminal Slepian-Wolf [0] (SW) and Wyner-Ziv [0] (WZ) theorems. Let us consider two binary correlated sources X and Y. If the two coders communicate, it is well known from Shannon's theory that the minimum lossless rate for X and Y is given by the joint entropy $H(X,Y)$. Slepian and Wolf have established in 1973 that this lossless compression rate bound can be approached with a vanishing error probability for long sequences, even if the two sources are coded separately, provided that they are decoded jointly and that their correlation is known to both the encoder and the decoder.

In 1976, Wyner and Ziv considered the problem of coding of two correlated sources X and Y, with respect to a fidelity criterion. They have established the rate-distortion function $R*_{X|Y}(D)$ for the case where the side information Y is perfectly known to the decoder only. For a given target distortion D, $R*_{X|Y}(D)$ in general verifies $R_{X|Y}(D) \leq R*_{X|Y}(D) \leq R_X(D)$, where $R_{X|Y}(D)$ is the rate required to encode X if Y is available to both the encoder and the decoder, and R_X is the minimal rate for encoding X without SI. These results give achievable rate bounds, however the design of codes and practical solutions for compression and communication applications remain a widely open issue.

[0] T. M. Cover and J. A. Thomas, Elements of Information Theory, Second Edition, July 2006.

[0] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources." IEEE Transactions on Information Theory, 19(4), pp. 471-480, July 1973.

[0] A. Wyner and J. Ziv, "The rate-distortion function for source coding ith side information at the decoder." IEEE Transactions on Information Theory, pp. 1-10, January 1976.

SMIS Project-Team

3. Research Program

3.1. Embedded Data Management

The challenge tackled is this research action is twofold: (1) to design embedded database techniques matching the hardware constraints of (current and future) smart objects and (2) to set up co-design rules helping hardware manufacturers to calibrate their future platforms to match the requirements of data driven applications. While a large body of work has been conducted on data management techniques for high-end servers (storage, indexation and query optimization models minimizing the I/O bottleneck, parallel DBMS, main memory DBMS, etc.), less research efforts have been placed on embedded database techniques. Light versions of popular DBMS have been designed for powerful handheld devices; yet DBMS vendors have never addressed the complex problem of embedding database components into chips. Proposals dedicated to databases embedded on chip usually consider small databases, stored in the non-volatile memory of the microcontroller –hundreds of kilobytes– and rely on NOR Flash or EEPROM technologies. Conversely, SMIS is pioneering the combination of microcontrollers and NAND Flash constraints to manage Gigabyte(s) size embedded databases. We present below the positioning of SMIS with respect to international teams conducting research on topics which may be connected to the addressed problem, namely work on electronic stable storage, RAM consumption and specific hardware platforms.

Major database teams are investigating data management issues related to hardware advances (EPFL: A. Ailamaki, CWI: M. Kersten, U. Of Wisconsin: J. M. Patel, Columbia: K. Ross, UCSB: A. El Abbadi, IBM Almaden: C. Mohan, etc.). While there are obvious links with our research on embedded databases, these teams target high-end computers and do not consider highly constrained architectures with non traditional hardware resources balance. At the other extreme, sensors (ultra-light computing devices) are considered by several research teams (e.g., UC Berkeley: D. Culler, ITU: P. Bonnet, Johns Hopkins University: A. Terzis, MIT: S. Madden, etc.). The focus is on the processing of continuous streams of collected data. Although the devices we consider share some hardware constraints with sensors, the objectives of both environments strongly diverge in terms of data cardinality and complexity, query complexity and data confidentiality requirements. Several teams are looking at efficient indexes on flash (HP LABS: G. Graefe, U. Minnesota: B. Debnath, U. Massachusetts: Y. Diao, Microsoft: S. Nath, etc.). Some studies try to minimize the RAM consumption, but the considered RAM/stable storage ratio is quite large compared to the constraints of the embedded context. Finally, a large number of teams have focused on the impact of flash memory on database system design (we presented an exhaustive state of the art in a VLDB tutorial [7]). The work conducted in the SMIS team on bi-modal flash devices takes the opposite direction, proposing to influence the design of flash devices by the expression of database requirements instead of running after the constantly evolving flash device technology.

3.2. Access and Usage Control Models

Access control management has been deeply studied for decades. Different models have been proposed to declare and administer access control policies, like DAC, MAC, RBAC, TMAC, and OrBAC. While access control management is well established, new models are being defined to cope with privacy requirements. Privacy management distinguishes itself from traditional access control is the sense that the data to be protected is personal. Hence, the user's consent must be reflected in the access control policies, as well as the usage of the data, its collection rules and its retention period, which are principles safeguarded by law and must be controlled carefully.

The research community working on privacy models is broad, and involves many teams worldwide including in France ENST-B, LIRIS, Inria LICIT, and LRI, and at the international level IBM Almaden, Purdue Univ., Politecnico di Milano and Univ. of Milano, George Mason Univ., Univ. of Massachusetts, Univ. of Texas and Colorado State Univ. to cite a few. Pioneer attempts towards privacy wary systems include the P3P Platform for Privacy Preservation [34] and Hippocratic databases [24]. In the last years, many other policy languages have been proposed for different application scenarios, including EPAL [38], XACML [36] and WSPL [29]. Hippocratic databases are inspired by the axiom that databases should be responsible for the privacy preservation of the data they manage. The architecture of a Hippocratic database is based on ten guiding principles derived from privacy laws.

The trend worldwide has been to propose enhanced access control policies to capture finer behavior and bridge the gap with privacy policies. To cite a few, Ardagna *et al.* (Univ. Milano) enables actions to be performed after data collection (like notification or removal), purpose binding features have been studied by Lefevre *et al.* (IBM Almaden), and Ni *et al.* (Purdue Univ.) have proposed obligations and have extended the widely used RBAC model to support privacy policies.

The positioning of the SMIS team within this broad area is rather (1) to focus on intuitive or automatic tools helping the individual to control some facets of her privacy (e.g., data retention, minimal collection) instead of increasing the expressiveness but also the complexity of privacy models and (2) to push concrete models enriched by real-case (e.g., medical) scenarios and by a joint work with researchers in Law.

3.3. Tamper-resistant Data Management

Tamper-resistance refers to the capacity of a system to defeat confidentiality and integrity attacks. This problem is complementary to access control management while being (mostly) orthogonal to the way access control policies are defined. Security surveys regularly point out the vulnerability of database servers against external (i.e., by intruders) and internal (i.e., by employees) attacks. Several attempts have been made in commercial DBMSs to strengthen server-based security, e.g., by separating the duty between DBA and DSA (Data Security Administrator), by encrypting the database footprint and by securing the cryptographic material using Hardware Security Modules (HSM) [31]. To face internal attacks, client-based security approaches have been investigated where the data is stored encrypted on the server and is decrypted only on the client side. Several contributions have been made in this direction, notably by U. of California Irvine (S. Mehrotra, Database Service Provider model), IBM Almaden (R. Agrawal, computation on encrypted data), U. of Milano (E. Damiani, encryption schemes), Purdue U. (E. Bertino, XML secure publication), U. of Washington (D. Suciu, provisional access) to cite a few seminal works. An alternative, recently promoted by Stony Brook Univ. (R. Sion), is to augment the security of the server by associating it with a tamper-resistant hardware module in charge of the security aspects. Contrary to traditional HSM, this module takes part in the query computation and performs all data decryption operations. SMIS investigates another direction based on the use of a tamper-resistant hardware module on the client side. Most of our contributions in this area are based on exploiting the tamper-resistance of secure tokens to build new data protection schemes.

While our work on Privacy-Preserving data Publishing (PPDP) is still related to tamper-resistance, a complementary positioning is required for this specific topic. The primary goal of PPDP is to anonymize/sanitize microdata sets before publishing them to serve statistical analysis purposes. PPDP (and privacy in databases in general) is a hot topic since 2000, when it was introduced by IBM Research (IBM Almaden: R. Agrawal, IBM Watson: C.C. Aggarwal), and many teams, mostly north American universities or research centres, study this topic (e.g., PORTIA DB-Privacy project regrouping universities such as Stanford with H. Garcia-Molina). Much effort has been devoted by the scientific community to the definition of privacy models exhibiting better privacy guarantees or better utility or a balance of both (such as differential privacy studied by C. Dwork: Microsoft Research or D. Kifer: Penn-State Univ and J. Gehrke: Cornell Univ) and thorough surveys exist that provide a large overview of existing PPDP models and mechanisms [35]. These works are however orthogonal to our approach in that they make the hypothesis of a trustworthy central server that can execute the anonymization process. In our work, this is not the case. We consider an architecture composed of a large

population of tamper-resistant devices weakly connected to an untrusted infrastructure and study how to compute PPDP problems in this context. Hence, our work has some connections with the works done on Privacy Preserving Data Collection (Stevens Institute of Tech. / Rutgers Univ,NJ: R.N.Wright, Univ Austin Texas: V. Shmatikov), on Secure Multi-party Computing for Privacy Preserving Data Mining (Rutgers Univ: J. Vaidya, Purdue Univ: C. Clifton) and on distributed PPDP algorithms (Univ Wisconsin: D. DeWitt, Univ Michigan: K. Lefevre, Rutgers Univ: J. Vaidya, Purdue Univ: C. Clifton) while none of them share the same architectural hypothesis as us.

STARS Project-Team

3. Research Program

3.1. Introduction

Stars follows three main research directions: perception for activity recognition, semantic activity recognition, and software engineering for activity recognition. **These three research directions are interleaved**: *the software engineering* reserach direction provides new methodologies for building safe activity recognition systems and *the perception* and *the semantic activity recognition* directions provide new activity recognition techniques which are designed and validated for concrete video analytics and healthcare applications. Conversely, these concrete systems raise new software issues that enrich the software engineering research direction.

Transversally, we consider a *new research axis in machine learning*, combining a priori knowledge and learning techniques, to set up the various models of an activity recognition system. A major objective is to automate model building or model enrichment at the perception level and at the understanding level.

3.2. Perception for Activity Recognition

Participants: Guillaume Charpiat, François Brémond, Sabine Moisan, Monique Thonnat.

Computer Vision; Cognitive Systems; Learning; Activity Recognition.

3.2.1. Introduction

Our main goal in perception is to develop vision algorithms able to address the large variety of conditions characterizing real world scenes in terms of sensor conditions, hardware requirements, lighting conditions, physical objects, and application objectives. We have also several issues related to perception which combine machine learning and perception techniques: learning people appearance, parameters for system control and shape statistics.

3.2.2. Appearance Models and People Tracking

An important issue is to detect in real-time physical objects from perceptual features and predefined 3D models. It requires finding a good balance between efficient methods and precise spatio-temporal models. Many improvements and analysis need to be performed in order to tackle the large range of people detection scenarios.

Appearance models. In particular, we study the temporal variation of the features characterizing the appearance of a human. This task could be achieved by clustering potential candidates depending on their position and their reliability. This task can provide any people tracking algorithms with reliable features allowing for instance to (1) better track people or their body parts during occlusion, or to (2) model people appearance for re-identification purposes in mono and multi-camera networks, which is still an open issue. The underlying challenge of the person re-identification problem arises from significant differences in illumination, pose and camera parameters. The re-identification approaches have two aspects: (1) establishing correspondences between body parts and (2) generating signatures that are invariant to different color responses. As we have already several descriptors which are color invariant, we now focus more on aligning two people detections and on finding their corresponding body parts. Having detected body parts, the approach can handle pose variations. Further, different body parts might have different influence on finding the correct match among a whole gallery dataset. Thus, the re-identification approaches have to search for matching strategies. As the results of the re-identification are always given as the ranking list, re-identification focuses on learning to rank. "Learning to rank" is a type of machine learning problem, in which the goal is to automatically construct a ranking model from a training data.

Therefore, we work on information fusion to handle perceptual features coming from various sensors (several cameras covering a large scale area or heterogeneous sensors capturing more or less precise and rich information). New 3D RGB-D sensors are also investigated, to help in getting an accurate segmentation for specific scene conditions.

Long term tracking. For activity recognition we need robust and coherent object tracking over long periods of time (often several hours in videosurveillance and several days in healthcare). To guarantee the long term coherence of tracked objects, spatio-temporal reasoning is required. Modelling and managing the uncertainty of these processes is also an open issue. In Stars we propose to add a reasoning layer to a classical Bayesian framework modelling the uncertainty of the tracked objects. This reasoning layer can take into account the a priori knowledge of the scene for outlier elimination and long-term coherency checking.

Controling system parameters. Another research direction is to manage a library of video processing programs. We are building a perception library by selecting robust algorithms for feature extraction, by insuring they work efficiently with real time constraints and by formalizing their conditions of use within a program supervision model. In the case of video cameras, at least two problems are still open: robust image segmentation and meaningful feature extraction. For these issues, we are developing new learning techniques.

3.2.3. Learning Shape and Motion

Another approach, to improve jointly segmentation and tracking, is to consider videos as 3D volumetric data and to search for trajectories of points that are statistically coherent both spatially and temporally. This point of view enables new kinds of statistical segmentation criteria and ways to learn them.

We are also using the shape statistics developed in [5] for the segmentation of images or videos with shape prior, by learning local segmentation criteria that are suitable for parts of shapes. This unifies patch-based detection methods and active-contour-based segmentation methods in a single framework. These shape statistics can be used also for a fine classification of postures and gestures, in order to extract more precise information from videos for further activity recognition. In particular, the notion of shape dynamics has to be studied.

More generally, to improve segmentation quality and speed, different optimization tools such as graph-cuts can be used, extended or improved.

3.3. Semantic Activity Recognition

Participants: Guillaume Charpiat, François Brémond, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding, Computer Vision

3.3.1. Introduction

Semantic activity recognition is a complex process where information is abstracted through four levels: signal (e.g. pixel, sound), perceptual features, physical objects and activities. The signal and the feature levels are characterized by strong noise, ambiguous, corrupted and missing data. The whole process of scene understanding consists in analyzing this information to bring forth pertinent insight of the scene and its dynamics while handling the low level noise. Moreover, to obtain a semantic abstraction, building activity models is a crucial point. A still open issue consists in determining whether these models should be given a priori or learned. Another challenge consists in organizing this knowledge in order to capitalize experience, share it with others and update it along with experimentation. To face this challenge, tools in knowledge engineering such as machine learning or ontology are needed.

Thus we work along the following research axes: high level understanding (to recognize the activities of physical objects based on high level activity models), learning (how to learn the models needed for activity recognition) and activity recognition and discrete event systems.

3.3.2. High Level Understanding

A challenging research axis is to recognize subjective activities of physical objects (i.e. human beings, animals, vehicles) based on a priori models and objective perceptual measures (e.g. robust and coherent object tracks).

To reach this goal, we have defined original activity recognition algorithms and activity models. Activity recognition algorithms include the computation of spatio-temporal relationships between physical objects. All the possible relationships may correspond to activities of interest and all have to be explored in an efficient way. The variety of these activities, generally called video events, is huge and depends on their spatial and temporal granularity, on the number of physical objects involved in the events, and on the event complexity (number of components constituting the event).

Concerning the modelling of activities, we are working towards two directions: the uncertainty management for representing probability distributions and knowledge acquisition facilities based on ontological engineering techniques. For the first direction, we are investigating classical statistical techniques and logical approaches. For the second direction, we built a language for video event modelling and a visual concept ontology (including color, texture and spatial concepts) to be extended with temporal concepts (motion, trajectories, events ...) and other perceptual concepts (physiological sensor concepts ...).

3.3.3. Learning for Activity Recognition

Given the difficulty of building an activity recognition system with a priori knowledge for a new application, we study how machine learning techniques can automate building or completing models at the perception level and at the understanding level.

At the understanding level, we are learning primitive event detectors. This can be done for example by learning visual concept detectors using SVMs (Support Vector Machines) with perceptual feature samples. An open question is how far can we go in weakly supervised learning for each type of perceptual concept (i.e. leveraging the human annotation task). A second direction is to learn typical composite event models for frequent activities using trajectory clustering or data mining techniques. We name composite event a particular combination of several primitive events.

3.3.4. Activity Recognition and Discrete Event Systems

The previous research axes are unavoidable to cope with the semantic interpretations. However they tend to let aside the pure event driven aspects of scenario recognition. These aspects have been studied for a long time at a theoretical level and led to methods and tools that may bring extra value to activity recognition, the most important being the possibility of formal analysis, verification and validation.

We have thus started to specify a formal model to define, analyze, simulate, and prove scenarios. This model deals with both absolute time (to be realistic and efficient in the analysis phase) and logical time (to benefit from well-known mathematical models providing re-usability, easy extension, and verification). Our purpose is to offer a generic tool to express and recognize activities associated with a concrete language to specify activities in the form of a set of scenarios with temporal constraints. The theoretical foundations and the tools being shared with Software Engineering aspects, they will be detailed in section 3.4 .

The results of the research performed in perception and semantic activity recognition (first and second research directions) produce new techniques for scene understanding and contribute to specify the needs for new software architectures (third research direction).

3.4. Software Engineering for Activity Recognition

Participants: Sabine Moisan, Annie Ressouche, Jean-Paul Rigault, François Brémond.

Software Engineering, Generic Components, Knowledge-based Systems, Software Component Platform, Object-oriented Frameworks, Software Reuse, Model-driven Engineering

The aim of this research axis is to build general solutions and tools to develop systems dedicated to activity recognition. For this, we rely on state-of-the art Software Engineering practices to ensure both sound design and easy use, providing genericity, modularity, adaptability, reusability, extensibility, dependability, and maintainability.

This research requires theoretical studies combined with validation based on concrete experiments conducted in Stars. We work on the following three research axes: *models* (adapted to the activity recognition domain), *platform architecture* (to cope with deployment constraints and run time adaptation), and *system verification* (to generate dependable systems). For all these tasks we follow state of the art Software Engineering practices and, if needed, we attempt to set up new ones.

3.4.1. Platform Architecture for Activity Recognition

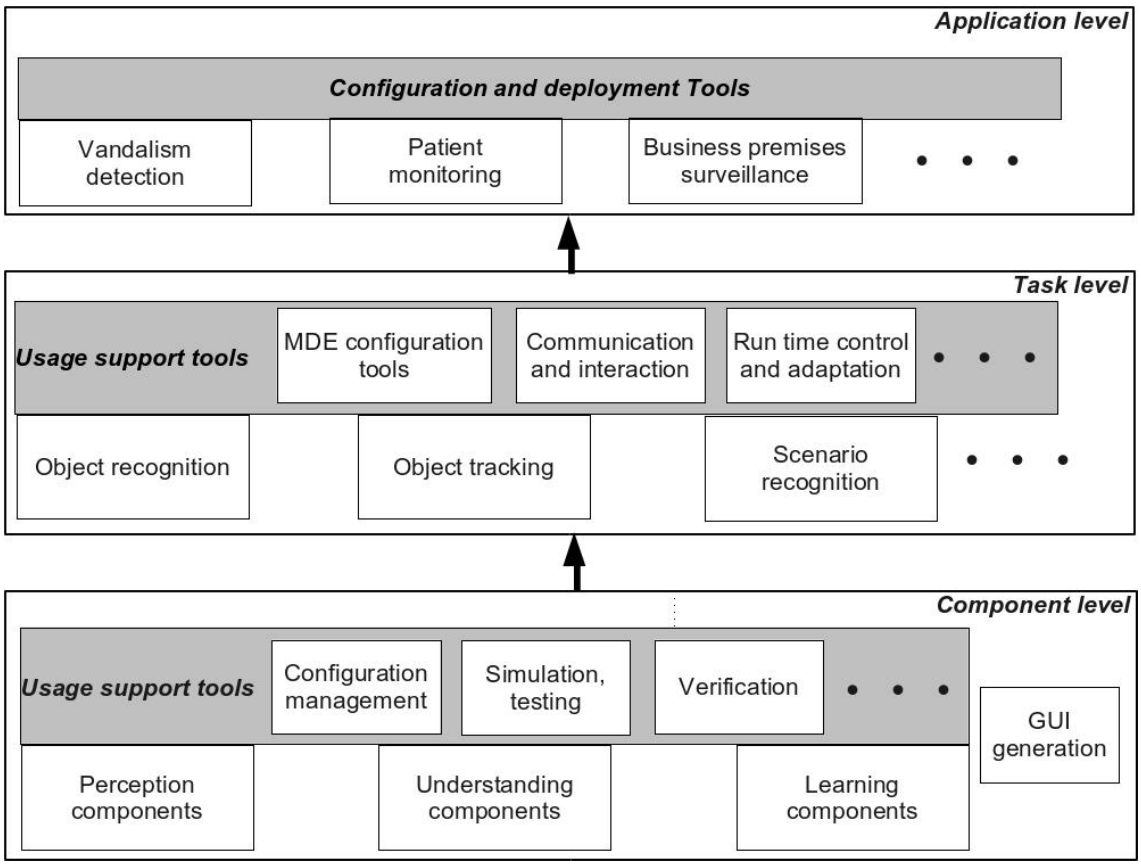


Figure 4. Global Architecture of an Activity Recognition The grey areas contain software engineering support modules whereas the other modules correspond to software components (at Task and Component levels) or to generated systems (at Application level).

In the former project teams Orion and Pulsar, we have developed two platforms, one (VSIP), a library of real-time video understanding modules and another one, LAMA [14], a software platform enabling to design not only knowledge bases, but also inference engines, and additional tools. LAMA offers toolkits to build and to adapt all the software elements that compose a knowledge-based system.

Figure 4 presents our conceptual vision for the architecture of an activity recognition platform. It consists of three levels:

- The **Component Level**, the lowest one, offers software components providing elementary operations and data for perception, understanding, and learning.

  - *Perception components* contain algorithms for sensor management, image and signal analysis, image and video processing (segmentation, tracking...), etc.
  - *Understanding components* provide the building blocks for Knowledge-based Systems: knowledge representation and management, elements for controlling inference engine strategies, etc.
  - *Learning components* implement different learning strategies, such as Support Vector Machines (SVM), Case-based Learning (CBL), clustering, etc.

  An Activity Recognition system is likely to pick components from these three packages. Hence, tools must be provided to configure (select, assemble), simulate, verify the resulting component combination. Other support tools may help to generate task or application dedicated languages or graphic interfaces.

- The **Task Level**, the middle one, contains executable realizations of individual tasks that will collaborate in a particular final application. Of course, the code of these tasks is built on top of the components from the previous level. We have already identified several of these important tasks: Object Recognition, Tracking, Scenario Recognition... In the future, other tasks will probably enrich this level.

  For these tasks to nicely collaborate, communication and interaction facilities are needed. We shall also add MDE-enhanced tools for configuration and run-time adaptation.

- The **Application Level** integrates several of these tasks to build a system for a particular type of application, e.g., vandalism detection, patient monitoring, aircraft loading/unloading surveillance, etc.. Each system is parameterized to adapt to its local environment (number, type, location of sensors, scene geometry, visual parameters, number of objects of interest...). Thus configuration and deployment facilities are required.

The philosophy of this architecture is to offer at each level a balance between the widest possible genericity and the maximum effective reusability, in particular at the code level.

To cope with real application requirements, we shall also investigate distributed architecture, real time implementation, and user interfaces.

Concerning implementation issues, we shall use when possible existing open standard tools such as NuSMV for model-checking, Eclipse for graphic interfaces or model engineering support, Alloy for constraint representation and SAT solving for verification, etc. Note that, in Figure 4 , some of the boxes can be naturally adapted from SUP existing elements (many perception and understanding components, program supervision, scenario recognition...) whereas others are to be developed, completely or partially (learning components, most support and configuration tools).

3.4.2. Discrete Event Models of Activities

As mentioned in the previous section (3.3) we have started to specify a formal model of scenario dealing with both absolute time and logical time. Our scenario and time models as well as the platform verification tools rely on a formal basis, namely the synchronous paradigm. To recognize scenarios, we consider activity descriptions as synchronous reactive systems and we apply general modelling methods to express scenario behaviour.

Activity recognition systems usually exhibit many safeness issues. From the software engineering point of view we only consider software security. Our previous work on verification and validation has to be pursued; in particular, we need to test its scalability and to develop associated tools. Model-checking is an appealing technique since it can be automatized and helps to produce a code that has been formally proved. Our verification method follows a compositional approach, a well-known way to cope with scalability problems in model-checking.

Moreover, recognizing real scenarios is not a purely deterministic process. Sensor performance, precision of image analysis, scenario descriptions may induce various kinds of uncertainty. While taking into account this uncertainty, we should still keep our model of time deterministic, modular, and formally verifiable. To formally describe probabilistic timed systems, the most popular approach involves probabilistic extension of timed automata. New model checking techniques can be used as verification means, but relying on model checking techniques is not sufficient. Model checking is a powerful tool to prove decidable properties but introducing uncertainty may lead to infinite state or even undecidable properties. Thus model checking validation has to be completed with non exhaustive methods such as abstract interpretation.

3.4.3. Model-Driven Engineering for Configuration and Control and Control of Video Surveillance systems

Model-driven engineering techniques can support the configuration and dynamic adaptation of video surveillance systems designed with our SUP activity recognition platform. The challenge is to cope with the many—functional as well as nonfunctional—causes of variability both in the video application specification and in the concrete SUP implementation. We have used *feature models* to define two models: a generic model of video surveillance applications and a model of configuration for SUP components and chains. Both of them express variability factors. Ultimately, we wish to automatically generate a SUP component assembly from an application specification, using models to represent transformations [56]. Our models are enriched with intra- and inter-models constraints. Inter-models constraints specify models to represent transformations. Feature models are appropriate to describe variants; they are simple enough for video surveillance experts to express their requirements. Yet, they are powerful enough to be liable to static analysis [75]. In particular, the constraints can be analysed as a SAT problem.

An additional challenge is to manage the possible run-time changes of implementation due to context variations (e.g., lighting conditions, changes in the reference scene, etc.). Video surveillance systems have to dynamically adapt to a changing environment. The use of models at run-time is a solution. We are defining adaptation rules corresponding to the dependency constraints between specification elements in one model and software variants in the other [55], [84], [78].

TITANE Project-Team

3. Research Program

3.1. Context

Geometric modeling and processing revolve around three main end goals: a computerized shape representation that can be visualized (creating a realistic or artistic depiction), simulated (anticipating the real) or realized (manufacturing a conceptual or engineering design). Aside from the mere editing of geometry, central research themes in geometric modeling involve conversions between physical (real), discrete (digital), and mathematical (abstract) representations. Going from physical to digital is referred to as shape acquisition and reconstruction; going from mathematical to discrete is referred to as shape approximation and mesh generation; going from discrete to physical is referred to as shape rationalization.

Geometric modeling has become an indispensable component for computational and reverse engineering. Simulations are now routinely performed on complex shapes issued not only from computer-aided design but also from an increasing amount of available measurements. The scale of acquired data is quickly growing: we no longer deal exclusively with individual shapes, but with entire *scenes*, possibly at the scale of entire cities, with many objects defined as structured shapes. We are witnessing a rapid evolution of the acquisition paradigms with an increasing variety of sensors and the development of community data, as well as disseminated data.

In recent years, the evolution of acquisition technologies and methods has translated in an increasing overlap of algorithms and data in the computer vision, image processing, and computer graphics communities. Beyond the rapid increase of resolution through technological advances of sensors and methods for mosaicing images, the line between laser scan data and photos is getting thinner. Combining, e.g., laser scanners with panoramic cameras leads to massive 3D point sets with color attributes. In addition, it is now possible to generate dense point sets not just from laser scanners but also from photogrammetry techniques when using a well-designed acquisition protocol. Depth cameras are getting increasingly common, and beyond retrieving depth information we can enrich the main acquisition systems with additional hardware to measure geometric information about the sensor and improve data registration: e.g., accelerometers or GPS for geographic location, and compasses or gyrometers for orientation. Finally, complex scenes can be observed at different scales ranging from satellite to pedestrian through aerial levels.

These evolutions allow practitioners to measure urban scenes at resolutions that were until now possible only at the scale of individual shapes. The related scientific challenge is however more than just dealing with massive data sets coming from increase of resolution, as complex scenes are composed of multiple objects with structural relationships. The latter relate i) to the way the individual shapes are grouped to form objects, object classes or hierarchies, ii) to geometry when dealing with similarity, regularity, parallelism or symmetry, and iii) to domain-specific semantic considerations. Beyond reconstruction and approximation, consolidation and synthesis of complex scenes require rich structural relationships.

The problems arising from these evolutions suggest that the strengths of geometry and images may be combined in the form of new methodological solutions such as photo-consistent reconstruction. In addition, the process of measuring the geometry of sensors (through gyrometers and accelerometers) often requires both geometry process and image analysis for improved accuracy and robustness. Modeling urban scenes from measurements illustrates this growing synergy, and it has become a central concern for a variety of applications ranging from urban planning to simulation through rendering and special effects.

3.2. Analysis

Complex scenes are usually composed of a large number of objects which may significantly differ in terms of complexity, diversity, and density. These objects must be identified and their structural relationships must be recovered in order to model the scenes with improved robustness, low complexity, variable levels of details and ultimately, semantization (automated process of increasing degree of semantic content).

Object classification is an ill-posed task in which the objects composing a scene are detected and recognized with respect to predefined classes, the objective going beyond scene segmentation. The high variability in each class may explain the success of the stochastic approach which is able to model widely variable classes. As it requires a priori knowledge this process is often domain-specific such as for urban scenes where we wish to distinguish between instances as ground, vegetation and buildings. Additional challenges arise when each class must be refined, such as roof super-structures for urban reconstruction.

Structure extraction consists in recovering structural relationships between objects or parts of object. The structure may be related to adjacencies between objects, hierarchical decomposition, singularities or canonical geometric relationships. It is crucial for effective geometric modeling through levels of details or hierarchical multiresolution modeling. Ideally we wish to learn the structural rules that govern the physical scene manufacturing. Understanding the main canonical geometric relationships between object parts involves detecting regular structures and equivalences under certain transformations such as parallelism, orthogonality and symmetry. Identifying structural and geometric repetitions or symmetries is relevant for dealing with missing data during data consolidation.

Data consolidation is a problem of growing interest for practitioners, with the increase of heterogeneous and defect-laden data. To be exploitable, such defect-laden data must be consolidated by improving the data sampling quality and by reinforcing the geometrical and structural relations sub-tending the observed scenes. Enforcing canonical geometric relationships such as local coplanarity or orthogonality is relevant for registration of heterogeneous or redundant data, as well as for improving the robustness of the reconstruction process.

3.3. Approximation

Our objective is to explore the approximation of complex shapes and scenes with surface and volume meshes, as well as on surface and domain tiling. A general way to state the shape approximation problem is to say that we search for the shape discretization (possibly with several levels of detail) that realizes the best complexity / distortion trade-off. Such problem statement requires defining a discretization model, an error metric to measure distortion as well as a way to measure complexity. The latter is most commonly expressed in number of polygon primitives, but other measures closer to information theory lead to measurements such as number of bits or minimum description length.

For surface meshes we intend to conceive methods which provide control and guarantees both over the global approximation error and over the validity of the embedding. In addition, we seek for resilience to heterogeneous data, and robustness to noise and outliers. This would allow repairing and simplifying triangle soups with cracks, self-intersections and gaps. Another exploratory objective is to deal generically with different error metrics such as the symmetric Hausdorff distance, or a Sobolev norm which mixes errors in geometry and normals.

For surface and domain tiling the term meshing is substituted for tiling to stress the fact that tiles may be not just simple elements, but can model complex smooth shapes such as bilinear quadrangles. Quadrangle surface tiling is central for the so-called *resurfacing* problem in reverse engineering: the goal is to tile an input raw surface geometry such that the union of the tiles approximates the input well and such that each tile matches certain properties related to its shape or its size. In addition, we may require parameterization domains with a simple structure. Our goal is to devise surface tiling algorithms that are both reliable and resilient to defect-laden inputs, effective from the shape approximation point of view, and with flexible control upon the structure of the tiling.

3.4. Reconstruction

Assuming a geometric dataset made out of points or slices, the process of shape reconstruction amounts to recovering a surface or a solid that matches these samples. This problem is inherently ill-posed as infinitely-many shapes may fit the data. One must thus regularize the problem and add priors such as simplicity or smoothness of the inferred shape.

The concept of geometric simplicity has led to a number of interpolating techniques commonly based upon the Delaunay triangulation. The concept of smoothness has led to a number of approximating techniques that commonly compute an implicit function such that one of its isosurfaces approximates the inferred surface. Reconstruction algorithms can also use an explicit set of prior shapes for inference by assuming that the observed data can be described by these predefined prior shapes. One key lesson learned in the shape problem is that there is probably not a single solution which can solve all cases, each of them coming with its own distinctive features. In addition, some data sets such as point sets acquired on urban scenes are very domain-specific and require a dedicated line of research.

In recent years the *smooth, closed case* (i.e., shapes without sharp features nor boundaries) has received considerable attention. However, the state-of-the-art methods have several shortcomings: in addition to being in general not robust to outliers and not sufficiently robust to noise, they often require additional attributes as input, such as lines of sight or oriented normals. We wish to devise shape reconstruction methods which are both geometrically and topologically accurate without requiring additional attributes, while exhibiting resilience to defect-laden inputs. Resilience formally translates into stability with respect to noise and outliers. Correctness of the reconstruction translates into convergence in geometry and (stable parts of) topology of the reconstruction with respect to the inferred shape known through measurements.

Moving from the smooth, closed case to the *piecewise smooth case* (possibly with boundaries) is considerably harder as the ill-posedness of the problem applies to each sub-feature of the inferred shape. Further, very few approaches tackle the combined issue of robustness (to sampling defects, noise and outliers) and feature reconstruction.

WILLOW Project-Team

3. Research Program

3.1. 3D object and scene modeling, analysis, and retrieval

This part of our research focuses on geometric models of specific 3D objects at the local (differential) and global levels, physical and statistical models of materials and illumination patterns, and modeling and retrieval of objects and scenes in large image collections. Our past work in these areas includes research aimed at recognizing rigid 3D objects in cluttered photographs taken from arbitrary viewpoints (Rothganger *et al.*, 2006), segmenting video sequences into parts corresponding to rigid scene components before recognizing these in new video clips (Rothganger *et al.*, 2007), retrieval of particular objects and buildings from images and videos (Sivic and Zisserman, 2003) and (Philbin *et al.*, 2007), and a theoretical study of a general formalism for modeling central and non-central cameras using the formalism and terminology of classical projective geometry (Ponce, 2009 and Batog *et al.*, 2010).

We have also developed multi-view stereopsis algorithms that have proven remarkably effective at recovering intricate details and thin features of compact objects and capturing the overall structure of large-scale, cluttered scenes. We have obtained a US patent 8,331,615 [0] for the corresponding software (PMVS, http://grail.cs.washington.edu/software/pmvs/) which is available under a GPL license and used for film production by ILM and Weta as well as by Google in Google Maps. It is also the basic technology used by Iconem, a start-up founded by Y. Ubelmann, a Willow collaborator. We have also applied our multi-view-stereo approach to model archaeological sites together with developing representations and efficient retrieval techniques to enable matching historical paintings to 3D models of archaeological sites (Russel *et al.*, 2011). Our current efforts in this area, outlined in detail in Section 6.2 , are focused on: (i) developing new representations of 3D architectural sites for matching and retrieval, (ii) modeling and recognition of objects in complex scenes using underlying 3D object models, and (iii) continuing our theoretical study of multi-view camera geometry.

3.2. Category-level object and scene recognition

The objective in this core part of our research is to learn and recognize quickly and accurately thousands of visual categories, including materials, objects, scenes, and broad classes of temporal events, such as patterns of human activities in picnics, conversations, etc. The current paradigm in the vision community is to model/learn one object category (read 2D aspect) at a time. If we are to achieve our goal, we have to break away from this paradigm, and develop models that account for the tremendous variability in object and scene appearance due to texture, material, viewpoint, and illumination changes within each object category, as well as the complex and evolving relationships between scene elements during the course of normal human activities.

Our current work, outlined in detail in Section 6.3 , has focused on: (i) capturing the spatial layout of objects using the formalism of graph matching, (ii) transferring mid-level image representations using convolutional neural networks, and (iii) learning the appearance of objects and their parts in a weakly supervised manner.

3.3. Image restoration, manipulation and enhancement

The goal of this part of our research is to develop models, and methods for image/video restoration, manipulation and enhancement. The ability to “intelligently" manipulate the content of images and video is just as essential as high-level content interpretation in many applications: This ranges from restoring old films or removing unwanted wires and rigs from new ones in post production, to cleaning up a shot of your daughter at her birthday party, which is lovely but noisy and blurry because the lights were out when she blew the candles, or editing out a tourist from your Roman holiday video. Going beyond the modest abilities of current “digital zoom" (bicubic interpolation in general) so you can close in on that birthday cake, “deblock" a football game on TV, or turn your favorite DVD into a blue-ray, is just as important.

[0] The patent: "Match, Expand, and Filter Technique for Multi-View Stereopsis" was issued December 11, 2012 and assigned patent number 8,331,615.

In this context, we believe there is a new convergence between computer vision, machine learning, and signal processing. For example: The idea of exploiting self-similarities in image analysis, originally introduced in computer vision for texture synthesis applications (Efros and Leung, 1999), is the basis for non-local means (Buades *et al.*, 2005), one of today's most successful approaches to image restoration. In turn, by combining a powerful sparse coding approach to non-local means (Dabov *et al.*, 2007) with modern machine learning techniques for dictionary learning (Mairal *et al.*, 2010), we have obtained denoising and demosaicking results that are the state of the art on standard benchmarks (Mairal *et al.*, 2009).

Our current work, outlined in detail in Section 6.4 , has focused on (i) image editing using accelerated local Laplacian filters and (ii) developing new formulation for image deblurring cast as a deep learning problem.

3.4. Human activity capture and classification

From a scientific point of view, visual action understanding is a computer vision problem that until recently has received little attention outside of extremely specific contexts such as surveillance or sports. Many of the current approaches to the visual interpretation of human activities are designed for a limited range of operating conditions, such as static cameras, fixed scenes, or restricted actions. The objective of this part of our project is to attack the much more challenging problem of understanding actions and interactions in unconstrained video depicting everyday human activities such as in sitcoms, feature films, or news segments. The recent emergence of automated annotation tools for this type of video data (Everingham, Sivic, Zisserman, 2006; Laptev, Marszałek, Schmid, Rozenfeld, 2008; Duchenne, Laptev, Sivic, Bach, Ponce, 2009) means that massive amounts of labelled data for training and recognizing action models will at long last be available. Our research agenda in this scientific domain is described below and our recent results are outlined in detail in Section 6.5 .

3.4.1. Weakly-supervised learning and annotation of human actions in video

We aim to leverage the huge amount of video data using readily-available annotations in the form of video scripts. Scripts, however, often provide only imprecise and incomplete information about the video. We address this problem with weakly-supervised learning techniques both at the text and image levels. To this end we recently explored automatic mining of scene and action categories. Within the PhD of Piotr Bojanowski we are currently extending this work towards exploiting richer textual descriptions of human actions and using them for learning more powerful contextual models of human actions in video.

3.4.2. Descriptors for video representation

Video representation has a crucial role for recognizing human actions and other components of a visual scene. Our work in this domain aims to develop generic methods for representing video data based on realistic assumptions. We explore the ways of enriching standard bag-of-feature representations with the higher-level information on objects, scenes and primitive human actions pre-learned on related tasks. We also investigate highly-efficient methods for computing video features motivated by the need of processing very large and increasing amounts of video.

3.4.3. Crowd characterization in video

Human crowds are characterized by distinct visual appearance and require appropriate tools for their analysis. In our work we develop generic methods for crowd analysis in video aiming to address multiple tasks such as (i) crowd density estimation and localization, (ii) characterization and recognition of crowd behaviours (e.g a person running against the crowd flow) as well as (iii) detection and tracking of individual people in the crowd. We address the challenge of analyzing crowds under the large variation in crowd density, video resolution and scene structure.

WIMMICS Project-Team

3. Research Program

3.1. Analyzing and Modeling Users, Communities and their Interactions in a Social Semantic Web Context

We rely on cognitive studies to build models of the system, the user and the interactions between users through the system, in order to support and improve these interactions.

In the short term, following the user modeling technique known as *Personas*, we are interested in these user models that are represented as specific, individual humans. *Personas* are derived from significant behavior patterns (i.e., sets of behavioral variables) elicited from interviews with and observations of users (and sometimes customers) of the future product. Our user models will specialize *Personas* approaches to include aspects appropriate to Web applications. The formalization of these models will rely on ontology-based modeling of users and communities starting with generalist schemas (e.g. FOAF: *Friend of a Friend*). In a longer term we will consider additional extensions of these schemas to capture additional aspects (e.g. emotional states). We will extend current descriptions of relational and emotional aspects in existing variants of the *Personas* technique.

Beyond the individual user models, we propose to rely on social studies to build models of the communities, their vocabularies, activities and protocols in order to identify where and when formal semantics is useful. In the short term we will further develop our method for elaborating collective personas and compare it to the related *collaboration personas* method and to the group modeling methods which are extensions to groups of the classical user modeling techniques dedicated to individuals. We also propose to rely on and adapt participatory sketching and prototyping to support the design of interfaces for visualizing and manipulating representations of collectives. In a longer term we want to focus on studying and modeling mixed representations containing social semantic representations (e.g. folksonomies) and formal semantic representations (e.g. ontologies) and propose operations that allow us to couple them and exchange knowledge between them.

Since we have a background in requirement models, we want to consider in the short term their formalization too in order to support mutual understanding and interoperability between requirements expressed with these heterogeneous models. In a longer term, we believe that argumentation theory can be combined to requirement engineering to improve participant awareness and support decision-making. On the methodological side, we propose to adapt to the design of such systems the incremental formalization approach originally introduced in the context of CSCW (Computer Supported Cooperative Work) and HCI (Human Computer Interaction) communities.

Finally, in the short term, for all the models we identified here we will rely on and evaluate knowledge representation methodologies and theories, in particular ontology-based modeling. In a longer term, additional models of the contexts, devices, processes and mediums will also be formalized and used to support adaptation, proof and explanation and foster acceptation and trust from the users. We specifically target a unified formalization of these contextual aspects to be able to integrate them at any stage of the processing.

3.2. Formalizing and Reasoning on Heterogeneous Semantic Graphs

Our second line of work is to formalize as typed graphs the models identified in the previous section in order to exploit them, e.g. in software. The challenge then is two-sided:

- To propose models and formalisms to capture and merge representations of both kinds of semantics (e.g. formal ontologies and social folksonomies). The important point is to allow us to capture those structures precisely and flexibly and yet create as many links as possible between these different objects.

- To propose algorithms (in particular graph-based reasoning) and approaches (e.g. human-computing methods) to process these mixed representations. In particular we are interested in allowing cross-enrichment between them and in exploiting the life cycle and specificities of each one to foster the life-cycles of the others.

While some of these problems are known, for instance in the field of knowledge representation and acquisition (e.g. disambiguation, fuzzy representations, argumentation theory), the Web reopens them with exacerbated difficulties of scale, speed, heterogeneity, and an open-world assumption.

Many approaches emphasize the logical aspect of the problem especially because logics are close to computer languages. We defend that the graph nature of Linked Data on the Web and the large variety of types of links that compose them call for typed graphs models. We believe the relational dimension is of paramount importance in these representations and we propose to consider all these representations as fragments of a typed graph formalism directly built above the Semantic Web formalisms. Our choice of a graph based programming approach for the semantic and social Web and of a focus on one graph based formalism is also an efficient way to support interoperability, genericity, uniformity and reuse.

ZENITH Project-Team

3. Research Program

3.1. Data Management

Data management is concerned with the storage, organization, retrieval and manipulation of data of all kinds, from small and simple to very large and complex. It has become a major domain of computer science, with a large international research community and a strong industry. Continuous technology transfer from research to industry has led to the development of powerful DBMSs, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (application servers, document systems, search engines, directories, etc.).

The fundamental principle behind data management is *data independence*, which enables applications and users to deal with the data at a high conceptual level while ignoring implementation details. The relational model, by resting on a strong theory (set theory and first-order logic) to provide data independence, has revolutionized data management. The major innovation of relational DBMS has been to allow data manipulation through queries expressed in a high-level (declarative) language such as SQL. Queries can then be automatically translated into optimized query plans that take advantage of underlying access methods and indices. Many other advanced capabilities have been made possible by data independence : data and metadata modeling, schema management, consistency through integrity rules and triggers, transaction support, etc.

This data independence principle has also enabled DBMS to continuously integrate new advanced capabilities such as object and XML support and to adapt to all kinds of hardware/software platforms from very small smart devices (smart phone, PDA, smart card, etc.) to very large computers (multiprocessor, cluster, etc.) in distributed environments.

Following the invention of the relational model, research in data management has continued with the elaboration of strong database theory (query languages, schema normalization, complexity of data management algorithms, transaction theory, etc.) and the design and implementation of DBMS. For a long time, the focus was on providing advanced database capabilities with good performance, for both transaction processing and decision support applications. And the main objective was to support all these capabilities within a single DBMS.

The problems of scientific data management (massive scale, complexity and heterogeneity) go well beyond the traditional context of DBMS. To address them, we capitalize on scientific foundations in closely related domains: distributed data management, cloud data management, big data, uncertain data management, metadata integration, data mining and content-based information retrieval.

3.2. Distributed Data Management

To deal with the massive scale of scientific data, we exploit large-scale distributed systems, with the objective of making distribution transparent to the users and applications. Thus, we capitalize on the principles of large-scale distributed systems such as clusters, peer-to-peer (P2P) and cloud, to address issues in data integration, scientific workflows, recommendation, query processing and data analysis.

Data management in distributed systems has been traditionally achieved by distributed database systems which enable users to transparently access and update several databases in a network using a high-level query language (e.g. SQL) [15]. Transparency is achieved through a global schema which hides the local databases' heterogeneity. In its simplest form, a distributed database system is a centralized server that supports a global schema and implements distributed database techniques (query processing, transaction management, consistency management, etc.). This approach has proved effective for applications that can benefit from centralized control and full-fledge database capabilities, e.g. information systems. However, it cannot scale up to more than tens of databases. Data integration systems, e.g. price comparators such as KelKoo, extend the distributed database approach to access data sources on the Internet with a simpler query language in read-only mode.

Parallel database systems extend the distributed database approach to improve performance (transaction throughput or query response time) by exploiting database partitioning using a multiprocessor or cluster system. Although data integration systems and parallel database systems can scale up to hundreds of data sources or database partitions, they still rely on a centralized global schema and strong assumptions about the network.

Scientific workflow management systems (SWfMS) such as Kepler (http://kepler-project.org) and Taverna (http://www.taverna.org.uk) allow scientists to describe and execute complex scientific procedures and activities, by automating data derivation processes, and supporting various functions such as provenance management, queries, reuse, etc. Some workflow activities may access or produce huge amounts of distributed data and demand high performance computing (HPC) environments with highly distributed data sources and computing resources. However, combining SWfMS with HPC to improve throughput and performance remains a difficult challenge. In particular, existing workflow development and computing environments have limited support for data parallelism patterns. Such limitation makes complex the automation and ability to perform efficient parallel execution on large sets of data, which may significantly slow down the execution of a workflow.

In contrast, peer-to-peer (P2P) systems [11] adopt a completely decentralized approach to data sharing. By distributing data storage and processing across autonomous peers in the network, they can scale without the need for powerful servers. Popular examples of P2P systems such as Gnutella and BitTorrent have millions of users sharing petabytes of data over the Internet. Although very useful, these systems are quite simple (e.g. file sharing), support limited functions (e.g. keyword search) and use simple techniques (e.g. resource location by flooding) which have performance problems. To deal with the dynamic behavior of peers that can join and leave the system at any time, they rely on the fact that popular data get massively duplicated.

Initial research on P2P systems has focused on improving the performance of query routing in the unstructured systems which rely on flooding, whereby peers forward messages to their neighbors. This work led to structured solutions based on Distributed Hash Tables (DHT), e.g. CHORD and Pastry, or hybrid solutions with super-peers that index subsets of peers. Another approach is to exploit gossiping protocols, also known as epidemic protocols. Gossiping has been initially proposed to maintain the mutual consistency of replicated data by spreading replica updates to all nodes over the network. It has since been successfully used in P2P networks for data dissemination. Basic gossiping is simple. Each peer has a complete view of the network (i.e., a list of all peers' addresses) and chooses a node at random to spread the request. The main advantage of gossiping is robustness over node failures since, with very high probability, the request is eventually propagated to all nodes in the network. In large P2P networks, however, the basic gossiping model does not scale as maintaining the complete view of the network at each node would generate very heavy communication traffic. A solution to scalable gossiping is by having each peer with only a partial view of the network, e.g. a list of tens of neighbor peers. To gossip a request, a peer chooses at random a peer in its partial view to send it the request. In addition, the peers involved in a gossip exchange their partial views to reflect network changes in their own views. Thus, by continuously refreshing their partial views, nodes can self-organize into randomized overlays which scale up very well.

We claim that a P2P solution is the right solution to support the collaborative nature of scientific applications as it provides scalability, dynamicity, autonomy and decentralized control. Peers can be the participants or organizations involved in collaboration and may share data and applications while keeping full control over their (local) data sources.

But for very-large scale scientific data analysis or to execute very large data-intensive workflow activities (activities that manipulate huge amounts of data), we believe cloud computing (see next section), is the right approach as it can provide virtually infinite computing, storage and networking resources. However, current cloud architectures are proprietary, ad-hoc, and may deprive users of the control of their own data. Thus, we postulate that a hybrid P2P/cloud architecture is more appropriate for scientific data management, by combining the bests of both approaches. In particular, it will enable the clean integration of the users' own computational resources with different clouds.

3.3. Cloud Data Management

Cloud computing encompasses on demand, reliable services provided over the Internet (typically represented as a cloud) with easy access to virtually infinite computing, storage and networking resources. Through very simple Web interfaces and at small incremental cost, users can outsource complex tasks, such as data storage, system administration, or application deployment, to very large data centers operated by cloud providers. Thus, the complexity of managing the software/hardware infrastructure gets shifted from the users' organization to the cloud provider. From a technical point of view, the grand challenge is to support in a cost-effective way the very large scale of the infrastructure which has to manage lots of users and resources with high quality of service.

Cloud customers could move all or part of their information technology (IT) services to the cloud, with the following main benefits:

- **Cost.** The cost for the customer can be greatly reduced since the IT infrastructure does not need to be owned and managed; billing is only based only on resource consumption. For the cloud provider, using a consolidated infrastructure and sharing costs for multiple customers reduces the cost of ownership and operation.
- **Ease of access and use.** The cloud hides the complexity of the IT infrastructure and makes location and distribution transparent. Thus, customers can have access to IT services anytime, and from anywhere with an Internet connection.
- **Quality of Service (QoS).** The operation of the IT infrastructure by a specialized provider that has extensive experience in running very large infrastructures (including its own infrastructure) increases QoS.
- **Elasticity.** The ability to scale resources out, up and down dynamically to accommodate changing conditions is a major advantage. In particular, it makes it easy for customers to deal with sudden increases in loads by simply creating more virtual machines.

However, cloud computing has some drawbacks and not all applications are good candidates for being "cloudified". The major concern is w.r.t. data security and privacy, and trust in the provider (which may use no so trustful providers to operate). One earlier criticism of cloud computing was that customers get locked in proprietary clouds. It is true that most clouds are proprietary and there are no standards for cloud interoperability. But this is changing with open source cloud software such as Hadoop, an Apache project implementing Google's major cloud services such as Google File System and MapReduce, and Eucalyptus, an open source cloud software infrastructure, which are attracting much interest from research and industry.

There is much more variety in cloud data than in scientific data since there are many different kinds of customers (individuals, SME, large corporations, etc.). However, we can identify common features. Cloud data can be very large, unstructured (e.g. text-based) or semi-structured, and typically append-only (with rare updates). And cloud users and application developers may be in high numbers, but not DBMS experts.

3.4. Big Data

Big data has become a buzz word, with different meanings depending on your perspective, e.g. 100 terabytes is big for a transaction processing system, but small for a web search engine. It is also a moving target, as shown by two landmarks in DBMS products: the Teradata database machine in the 1980's and the Oracle Exadata database machine in 2010.

Although big data has been around for a long time, it is now more important than ever. We can see overwhelming amounts of data generated by all kinds of devices, networks and programs, e.g. sensors, mobile devices, internet, social networks, computer simulations, satellites, radiotelescopes, etc. Storage capacity has doubled every 3 years since 1980 with prices steadily going down (e.g. 1 Gigabyte for: 1M$ in 1982, 1K$ in 1995, 0.12$ in 2011), making it affordable to keep more data. And massive data can produce high-value information and knowledge, which is critical for data analysis, decision support, forecasting, business intelligence, research, (data-intensive) science, etc.

The problem of big data has three main dimensions, quoted as the three big V's:

- Volume: refers to massive amounts of data, making it hard to store, manage, and analyze (big analytics);
- Velocity: refers to continuous data streams being produced, making it hard to perform online processing and analysis;
- Variety: refers to different data formats, different semantics, uncertain data, multiscale data, etc., making it hard to integrate and analyze.

There are also other V's like: validity (is the data correct and accurate?); veracity (are the results meaningful?); volatility (how long do you need to store this data?).

Current big data management (NoSQL) solutions have been designed for the cloud, as cloud and big data are synergistic. They typically trade consistency for scalability, simplicity and flexibility. They use a radically different architecture than RDBMS, by exploiting (rather than embedding) a distributed file system such as Google File System (GFS) or Hadoop Distributed File System (HDFS), to store and manage data in a highly fault-tolerant manner. They tend to rely on a more specific data model, e.g. key-value store such as Google Bigtable, Hadoop Hbase or Apache CouchDB) with a simple set of operators easy to use from a programming language. For instance, to address the requirements of social network applications, new solutions rely on a graph data model and graph-based operators. User-defined functions also allow for more specific data processing. MapReduce is a good example of generic parallel data processing framework, on top of a distributed file system (GFS or HDFS). It supports a simple data model (sets of (key, value) pairs), which allows user-defined functions (map and reduce). Although quite successful among developers, it is relatively low-level and rigid, leading to custom user code that is hard to maintain and reuse. In Zenith, we exploit or extend MapReduce and NoSQL technologies to fit our needs for scientific workflow management and scalable data analysis.

3.5. Uncertain Data Management

Data uncertainty is present in many scientific applications. For instance, in the monitoring of plant contamination by INRA teams, sensors generate periodically data which may be uncertain. Instead of ignoring (or correcting) uncertainty, which may generate major errors, we need to manage it rigorously and provide support for querying.

To deal with uncertainty, there are several approaches, e.g. probabilistic, possibilistic, fuzzy logic, etc. The *probabilistic approach* is often used by scientists to model the behavior of their underlying environments. However, in many scientific applications, data management and uncertain query processing are not integrated, i.e., the queries are usually answered using ad-hoc methods after doing manual or semi-automatic statistical treatment on the data which are retrieved from a database. In Zenith, we aim at integrating scientific data management and query processing within one system. This should allow scientists to issue their queries in a query language without thinking about the probabilistic treatment which should be done in background in order to answer the queries. There are two important issues which any PDBMS should address: 1) how to represent a probabilistic database, i.e., data model; 2) how to answer queries using the chosen representation, i.e., query evaluation.

One of the problems on which we focus is *scalable query processing* over uncertain data. A naive solution for evaluating probabilistic queries is to enumerate all possible worlds, i.e., all possible instances of the database, execute the query in each world, and return the possible answers together with their cumulative probabilities. However, this solution can not scale up due to the exponential number of possible worlds which a probabilistic database may have. Thus, the problem is quite challenging, particularly due to the exponential number of possibilities that should be considered for evaluating queries. In addition, most of our underlying scientific applications are not centralized; the scientists share part of their data in a *P2P* manner. This distribution of data makes very complicated the processing of probabilistic queries. To develop efficient query processing techniques for distributed scientific applications, we can take advantage of two main distributed technologies: *P2P* and *Cloud*. Our research experience in P2P systems has proved us that we can propose scalable solutions

for many data management problems. In addition, we can use the cloud parallel solutions, e.g. MapReduce, to parallelize the task of query processing, when possible, and answer queries of scientists in reasonable execution times. Another challenge for supporting scientific applications is uncertain data integration. In addition to managing the uncertain data for each user, we need to integrate uncertain data from different sources. This requires revisiting traditional data integration in major ways and dealing with the problems of uncertain mediated schema generation and uncertain schema mapping.

3.6. Big data Integration

Nowadays, scientists can rely on web 2.0 tools to quickly share their data and/or knowledge (e.g. ontologies of the domain knowledge). Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). To make high numbers of scientific data sources easily accessible to community members, it is necessary to identify semantic correspondences between metadata structures or models of the related data sources. The main underlying task is called matching, which is the process of discovering semantic correspondences between metadata structures such as database schema and ontologies. Ontology is a formal and explicit description of a shared conceptualization in terms of concepts (i.e., classes, properties and relations). For example, the matching may be used to align gene ontologies or anatomical metadata structures.

To understand a data source content, metadata (data that describe the data) is crucial. Metadata can be initially provided by the data publisher to describe the data structure (e.g. schema), data semantics based on ontologies (that provide a formal representation of the domain knowledge) and other useful information about data provenance (publisher, tools, methods, etc.). Scientific metadata is very heterogeneous, in particular because of the great autonomy of the underlying data sources, which leads to a large variety of models and formats. The high heterogeneity makes the matching problem very challenging. Furthermore, the number of ontologies and their size grow fastly, and so does their diversity and heterogeneity. As a result, schema/ontology matching has become a prominent and challenging topic.

3.7. Data Mining

Data mining provides methods to discover new and useful patterns from very large sets of data. These patterns may take different forms, depending on the end-user's request, such as:

- **Frequent itemsets and association rules [1]**. In this case, the data is usually a table with a high number of rows and the algorithm extracts correlations between column values. This problem was first motivated by commercial and marketing purposes (*e.g.* discovering frequent correlations between items bought in a shop, which could help selling more). A typical example of frequent itemset from a sensor network in a smart building would say that "in 20% rooms, the door is closed, the room is empty, and lights are on."
- **Frequent sequential pattern extraction.** This problem is very similar to frequent itemset mining, but in this case, the order between events has to be considered. Let us consider the smart-building example again. A frequent sequence, in this case, could say that "in 40% rooms, lights are on at time i, the room is empty at time i+j and the door is closed at time i+j+k". Discovering frequent sequences has become a crucial need in marketing, but also in security (detecting network intrusions for instance) in usage analysis (web usage is one of the main applications) and any domain where data arrive in a specific order (usually given by timestamps).
- **Clustering [14]**. The goal of clustering algorithms is to group together data that have similar characteristics, while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we would find clusters of rooms, where offices will be in one category and copy machine rooms in another one because of their characteristics (hours of people presence, number of times lights are turned on and off, etc.).

One of the main problems for data mining methods has been to deal with data streams. Actually, data mining methods have first been designed for very large data sets where complex algorithms of artificial intelligence were not able to complete within reasonable time responses because of data size. The problem was thus to find a good trade-off between response time and results relevance. The patterns described above well match this trade-off since they both provide interesting knowledge for data analysts and allow algorithm having good time complexity on the number of records. Itemset mining algorithms, for instance, depend more on the number of columns (for a sensor it would be the number of possible items such as temperature, presence, status of lights, etc.) than the number of lines (number of sensors in the network). However, with the ever growing size of data and their production rate, a new kind of data source has recently emerged as data streams. A data stream is a sequence of events arriving at high rate. By "high rate", we usually admit that traditional data mining methods reach their limits and cannot complete in real-time, given the data size. In order to extract knowledge from such streams, a new trade-off had to be found and the data mining community has investigated approximation methods that could allow to maintain a good quality of results for the above patterns extraction.

For scientific data, data mining now has to deal with new and challenging characteristics. First, scientific data is often associated to a level of uncertainty (typically, sensed values have to be associated to the probability that this value is correct or not). Second, scientific data might be extremely large and need cloud computing solutions for their storage and analysis. Eventually, we will have to deal with high dimension and heterogeneous data.

3.8. Content-based Information Retrieval

Today's technologies for searching information in scientific data mainly rely on relational DBMS or text-based indexing methods. However, content-based information retrieval has progressed much in the last decade and is now considered as one of the most promising for future search engines. Rather than restricting search to the use of metadata, content-based methods attempt to index, search and browse digital objects by means of signatures describing their actual content. Such methods have been intensively studied in the multimedia community to allow searching the massive amount or raw multimedia documents created every day (e.g. 99% of web data are audio-visual content with very sparse metadata). Successful and scalable content-based methods have been proposed for searching objects in large image collections or detecting copies in huge video archives. Besides multimedia contents, content-based information retrieval methods recently started to be studied on more diverse data such as medical images, 3D models or even molecular data. Potential applications in scientific data management are numerous. First of all, to allow searching the huge collections of scientific images (earth observation, medical images, botanical images, biology images, etc.) but also to browse large datasets of experimental data (e.g. multisensor data, molecular data or instrumental data). Despite recent progress, scalability remains a major issue, involving complex algorithms (such as similarity search, clustering or supervised retrieval), in high dimensional spaces (up to millions of dimensions) with complex metrics (Lp, Kernels, sets intersections, edit distances, etc.). Most of these algorithms have linear, quadratic or even cubic complexities so that their use at large scale is not affordable without consistent breakthrough. In Zenith, we plan to investigate the following challenges:

- **High-dimensional similarity search**. Whereas many indexing methods were designed in the last 20 years to efficiently retrieve multidimensional data with relatively small dimensions, high-dimensional data have been more challenging due to the well-known dimensionality curse. Only recently have some methods appeared that allow approximate Nearest Neighbors queries in sub-linear time. In particular, Locality Sensitive Hashing methods which offer new theoretical insights in high-dimensional Euclidean spaces and proved the interest of random projections. But there are still some challenging issues that need to be solved including efficient similarity search in any kernel or metric spaces, efficient construction of knn-graphs or relational similarity queries.
- **Large-scale supervised retrieval**. Supervised retrieval aims at retrieving relevant objects in a dataset by providing some positive and/or negative training samples. To solve such a task, there has been a focused interest on using Support Vector Machines (SVM) that offer the possibility to construct generalized, non-linear predictors in high-dimensional spaces using small training

sets. The prediction time complexity of these methods is usually linear in dataset size. Allowing hyperplane similarity queries in sub-linear time is for example a challenging research issue. A symmetric problem in supervised retrieval consists in retrieving the most relevant object categories that might contain a given query object, providing huge labeled datasets (up to millions of classes and billions of objects) and very few objects per category (from 1 to 100 objects). SVM methods that are formulated as quadratic programming with cubic training time complexity and quadratic space complexity are clearly not usable. Promising solutions to such problems include hybrid supervised-unsupervised methods and supervised hashing methods.

- **Distributed content-based retrieval**. Distributed content-based retrieval methods appeared recently as a promising solution to manage masses of data distributed over large networks, particularly when the data cannot be centralized for privacy or cost reasons (which is often the case in scientific social networks, e.g. botanist social networks). However, current methods are limited to very simple similarity search paradigms. In Zenith, we will consider more advanced distributed content-based retrieval and mining methods such as k-nn graphs construction, large-scale supervised retrieval or multi-source clustering.