RESEARCH CENTER

FIELD

Activity Report 2012

Section Scientific Foundations

Edition: 2013-04-24

APPLIED MATHEMATICS, COMPUTATION AND SIMULATION

BYMOORE Exploratory Action (section vide)

POPIX Exploratory Action

3. Scientific Foundations

3.1. Scientific Foundations

Mathematical models that characterize comple 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 characterizeing 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
- mixed hidden Markov models
- mixture models and model mixtures
- time-to-event models
- models including a large number of covariates

ABSTRACTION Project-Team

3. Scientific Foundations

3.1. Abstract Interpretation Theory

The abstract interpretation theory [41], [32], [42], is the main scientific foundation of the work of the ABSTRACTION project-team. Its main current application is on the safety and security of complex hardware and software computer systems either sequential [41], [34], or parallel [36] with shared memory [33], [35], [44] or synchronous message [43] communication.

Abstract interpretation is a theory of sound approximation of mathematical structures, in particular those involved in the behavior of computer systems. It allows the systematic derivation of sound methods and algorithms for approximating undecidable or highly complex problems in various areas of computer science (semantics, verification and proof, model-checking, static analysis, program transformation and optimization, typing, software steganography, etc...) and system biology (pathways analysis).

3.2. Formal Verification by Abstract Interpretation

The *formal verification* of a program (and more generally a computer system) consists in proving that its *semantics* (describing "what the program executions actually do") satisfies its *specification* (describing "what the program executions are supposed to do").

Abstract interpretation formalizes the idea that this formal proof can be done at some level of abstraction where irrelevant details about the semantics and the specification are ignored. This amounts to proving that an *abstract semantics* satisfies an *abstract specification*. An example of abstract semantics is Hoare logic while examples of abstract specifications are invariance, partial, or total correctness. These examples abstract away from concrete properties such as execution times.

Abstractions should preferably be *sound* (no conclusion derived from the abstract semantics is wrong with respect to the program concrete semantics and specification). Otherwise stated, a proof that the abstract semantics satisfies the abstract specification should imply that the concrete semantics also satisfies the concrete specification. Hoare logic is a sound verification method, debugging is not (since some executions are left out), bounded model checking is not either (since parts of some executions are left out). Unsound abstractions lead to *false negatives* (the program may be claimed to be correct/non erroneous with respect to the specification whereas it is in fact incorrect). Abstract interpretation can be used to design sound semantics and formal verification methods (thus eliminating all false negatives).

Abstractions should also preferably be *complete* (no aspect of the semantics relevant to the specification is left out). So if the concrete semantics satisfies the concrete specification this should be provable in the abstract. However program proofs (for non-trivial program properties such as safety, liveness, or security) are undecidable. Nevertheless, we can design tools that address undecidable problems by allowing the tool not to terminate, to be driven by human intervention, to be unsound (e.g. debugging tools omit possible executions), or to be incomplete (e.g. static analysis tools may produce false alarms). Incomplete abstractions lead to *false positives* or *false alarms* (the specification is claimed to be potentially violated by some program executions while it is not). Semantics and formal verification methods designed by abstract interpretation may be complete (e.g. [39], [40], [47]) or incomplete (e.g. [2]).

Sound, automatic, terminating and precise tools are difficult to design. Complete automatic tools to solve non-trivial verification problems cannot exist, by undecidability. However static analysis tools producing very few or no false alarms have been designed and used in industrial contexts for specific families of properties and programs [45]. In all cases, abstract interpretation provides a systematic construction method based on the effective approximation of the concrete semantics, which can be (partly) automated and/or formally verified.

Abstract interpretation aims at:

- providing a basic coherent and conceptual theory for understanding in a unified framework the multiplicity of ideas, concepts, reasonings, methods, and tools on formal program analysis and verification [41], [42];
- guiding the correct formal design of *abstract semantics* [40], [47] and automatic tools for *program analysis* (computing an abstract semantics) and *program verification* (proving that an abstract semantics satisfies an abstract specification) [37].

Abstract interpretation theory studies semantics (formal models of computer systems), abstractions, their soundness, and completeness.

In practice, abstract interpretation is used to design analysis, compilation, optimization, and verification tools which must automatically and statically determine properties about the runtime behavior of programs. For example the ASTRÉE static analyzer (Section 5.2), which was developed by the team over the last decade, aims at proving the absence of runtime errors in programs written in the C programming language. It was originally used in the aerospace industry to verify very large, synchronous, time-triggered, real-time, safety-critical, embedded software and its scope of application was later broadly widened. ASTRÉE is now industrialized by AbsInt Angewandte Informatik GmbH and is commercially available.

3.3. Advanced Introductions to Abstract Interpretation

A short, informal, and intuitive introduction to the theory of abstract interpretation can be found in [37], see also "AbstractInterpretationinaNutshell" [1] on the web. A more comprehensive introduction is available online [2]. The paper entitled "Basicconceptsofabstractinterpretation" [38] and an elementary "courseonabstract interpretation" [3] can also be found on the web.

[1] www.di.ens.fr/~cousot/AI/IntroAbsInt.html
[2] www.di.ens.fr/~cousot/AI/
[3] web.mit.edu/afs/athena.mit.edu/course/16/16.399/www/

ALF Project-Team

3. Scientific Foundations

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- cores 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 which 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 [34] 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 [36]). Such simulations are CPU consuming and simulations cannot be run on a complete application. Sampling representative slices of the application was proposed [5] and popularized by the Simpoint [45] 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 [38]), accurate evaluation of the WCET of an algorithm on a complex uniprocessor system is a difficult problem. Accurately modelling data cache behavior [4] 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) [42].

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 [49].

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

The overall performance of a multicore system depends on many parameters including architecture, operating system, runtime environment, compiler technology and application development. Even the architecture dimension of a 1000's core system cannot be explored by a single research project. Many research groups are exploring the parallel dimension of the multicores essentially targeting issues such as coherency and scalability.

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 [14], [13]. 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 [37]. 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 [44], but the most promising solutions integrate hardware support for software thread-level speculation [47].

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* [10] through techniques such as clock gating or throttling and *activity migration* [46][7].

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 [9], 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 [39] 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 [40] 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 [3]. Moreover, this bytecode representation can be compiled to native code with performance similar to static compilation [2]. 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 maintenancesince 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 informations 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 onlyhigh 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. Considering the ALF base architecture, 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) [42]. 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 [41] 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 [50], therefore hardware-based or compiler methods such as [48] 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.

AOSTE Project-Team

3. Scientific Foundations

3.1. Models of Computation and Communication (MoCCs)

Participants: Charles André, Robert de Simone, 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 this 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: Charles André, 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, that result is a set of algorithm graphs, "architecture-aware", corresponding to refinements of the initial algorithm graph, by computing spatial (distribution) and timing (scheduling) allocations of the operations according to the architecture graph resource availability. In that context "Adequation" refers to some search amongst the solution space of resulting algorithm graphs, labelled by timing characteristics, for one 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: monoprocessor real-time scheduling 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 our monoprocessor real-time scheduling work, beside the classical deadline constraint, often equal to a period, we take into consideration dependences beetween tasks and several, possibly related, latencies. A latency is a generalization of the typical "end-to-end" constraint. Dealing with multiple real-time constraints raises the complexity of that issue. Moreover, because the preemption leads 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 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, but seeking the exact cost of the preemption such that it can be integrated in schedulability conditions, 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 exact cost of preemption is the most difficult part because it varies according to the instance of each task [10]. Finally, we investigate also the problem of mixing hard real-time and soft real-time constraints that arises in the most complex applications.

The second research area is devoted to distributed real-time scheduling with embedding constraints. We use the results obtained in the monoprocessor case in order to derive solutions for the problem of multiprocessor (distributed) real-time scheduling. In addition to satisfy the multiple real-time constraints mentioned in the monoprocessor case, we have to minimize the total execution time (makespan) since we deal with automatic control applications involving feedback. Furthermore, the domain of embedded systems leads to solving minimization resources problems. Since these optimization problems are of NP-hard complexity 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 such as variants of "simulated annealing".

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).

Since real-time distributed systems are often safety-critical we address dependability issues, to tolerate faults in processors and communication interconnects. We maily focus on software redondancy, rather than hardware, to ensure real-time behaviour preservation in presence of faulty processors and/or communication media (where possible failures are predictively specified by the designer). We investigate fail silent, transient, intermittent, and Byzantine faults.

ARIC Team

3. Scientific Foundations

3.1. Applications

Whether its purpose is to design better operators or to make the best use of existing ones, computer arithmetic is strongly connected to applications. Some application domains are particularly in demand for high-quality arithmetic: high-performance computing (HPC) for floating-point, accounting for decimal, digital signal processing (DSP) for fixed-point, embedded systems for application-specific operators, cryptography for finite fields. Each domain comes with its specific constraints and quality metric. For example, cryptography has a specific need of resistance to attacks that impact the design of the operators themselves: a good operator for cryptography should have electromagnetic emissions and power consumption patterns independent of the data it manipulates. Another example is very large-scale HPC, which in some cases is reaching the limits of the accuracy provided by the prevalent double-precision floating-point arithmetic.

The regional (Rhône-Alpes) context is especially strong in embedded systems, with the Minalogic Competitivity Centre, major players such as STMicroelectronics, CEA and Inria, and strong startups such as Kalray. This is also true at the European level, with the HiPEAC European network of excellence. This network addresses hardware issues, but also software and compiler issues.

Indeed, the bridge between the application and the underlying hardware arithmetic is usually the compiler. Therefore, more and more arithmetic expertise should be integrated within the compiler. This goes on par with the current trend to automate arithmetic core generation. In the long term, working at the compiler level opens optimization perspective beyond what compilers traditionally perform, for instance ad-hoc generation and optimization in context of application-specific functional cores.

However, much of computer arithmetic research still focuses on the implementation of standard computing cores (such as elementary functions, linear algebra operators, or DSP filters), although this implementation is more and more automated as illustrated by projects such as ATLAS, Spiral, FFTW, and others.

Cryptography is an active field of research where there is a strong demand for efficient arithmetic operators. Practical schemes such as hash functions, public-key encryption and digital signatures may be used in constrained environments, leading to interesting arithmetic problems. Common examples are long integer arithmetic (RSA) and arithmetic of algebraic curves and finite fields of medium sizes (elliptic curve cryptography, including pairing-based cryptography), and small finite fields (code-based cryptography and lattice-based cryptography).

3.2. Technology

The traditional arithmetic operators are small, low-level, close-to-the-silicon hardware building bricks, and it is therefore important to anticipate the evolutions of the technology to address the new challenges these evolutions will bring.

It is well known now that Moore's law is no longer what it used to be. It continues to bring more transistors on a chip with each new generation, but the speed of these transistors no longer increases, and their power consumption no longer decreases. With more integration come also more reliability issues.

These are the driving forces behind the shift to multicore processors, and to coarser and more complex processing units in these processors: single-instruction, multiple data (SIMD) instructions, fused multiply-and-add, and soon dot-product operations. It also led to the emergence of new massively parallel computing devices such as graphical processing units (GPU) and field-programmable gate arrays (FPGAs). Both are increasingly being used for general purpose computing.

In the shift to massively parallel multicores and GPUs, the real challenge is how to program them. With respect to computer arithmetic, the main problem is the control of numerical precision: the order of the elementary operations is changed in a parallel execution, and will very often not even be deterministic if the main objective is performance. Assessing or guaranteeing numerical quality in the face of this uncertainty is an open problem, all the more as SIMD units and limited data bandwidth encourage the use of mixed precision where possible.

Concerning FPGAs, their programming model is that of a digital circuit which may be application-specific, and even change in the lifetime of an application. The challenge here is to design arithmetic operators that exploit this reconfigurability, which is their main strength. Whereas processor operators have to be as general-purpose as possible, in an FPGA an operator can be designed specifically for a given application's context. A related challenge is to convince application designers that they should use these operators, which may be radically different from those they are used to see in processors. The C-to-hardware community addresses this challenge by hiding the FPGA behind a classical C programming model. This raises the arithmetic problem of automatically extracting from a piece of C code a fragment that is suitable for implementation as an application-specific operator in an FPGA.

In traditional circuit design, power consumption is no longer a concern only for embedded, battery-powered applications: heat dissipation is now the main issue limiting the frequency of high-performance processors. The nature of power consumption is also changing: it used to be caused mostly by the active switching transistors, but leakage power is now as much of a concern. All this impacts the design of operators, but also their use: the energy-per-computation metric will become more and more important and will orient algorithmic choices, for instance inviting us to reassess the benefits of pre-computing values.

Finally, the industry is preparing to address, within a decade or two, the end of silicon-based Moore's law. In addition to the physical limits (it is believed so far that we need at least one atom to build a transistor), the raising cost of fabrication plants at each generation has led to increasing concentration in fewer and fewer foundries. There will therefore be an economic limit when the number of foundries is down to one. Silicon replacement alternative are emerging in laboratories, without a clear winner yet. When these alternatives reach the integrated circuit, they may be expected to drastically change the rules by which arithmetic operators are designed.

3.3. Numbers and Number Representation

The first issue addressed by computer arithmetic is the representation of numbers in the computer. There are many possible representations, and a representation typically has many parameters. For instance, for integers, the decimal representation and the binary representation belong to the same family, only differing by the radix, 10 or 2. Another parameter of this representation is the number of digits considered.

A good representation is one that enables good computing. Here the measures of quality are numerous, sometimes conflicting, and application-dependent. For instance, the classical representation of integers is compact, but addition involves a carry propagation. There exists another classical family of integer representations which are redundant, therefore less compact, but allow for carry-free, thus faster, addition. Many other quality measures are possible, for instance power consumption, or silicon area.

Research on number representation for integers and reals is no longer very active, and it may be that there is little left to find in this field. The corresponding expertise now belongs to the common culture of the computer arithmetic community. For the integers, from time to time, a new context revives interest in an exotic number representation. For the reals, the indisputable advantages of a widespread and shared standard (the IEEE 754 floating-point standard) weigh strongly against innovation. However, for barely more complex datatypes, such as complex numbers or real intervals (each of which can be represented by a pair of reals), there is no such consensus yet.

Finally, research on number representation is still very active for datatypes related to more recent application fields, most notably in cryptography. For instance, the elliptic curve number system has been introduced because it allowed to use smaller keys for similar security, and research is still active to find representations of elliptic curves that enable efficient computation on this number system. This research tries to improve on

the usual quality metrics (performance, resource consumption, power), and in addition we have two more context-specific metrics: the key size, and the security level.

3.4. Arithmetic Algorithms

Each year, new algorithms are still published for basic operations (from addition to division), but the main focus of the computer arithmetic community has long shifted to more complex objects: examples are sums of many numbers, arithmetic on complex numbers, and evaluation of algebraic and transcendental functions.

The latter typically reduces to polynomial evaluation, with two sub-problems: firstly, one must find a good approximation polynomial. Secondly, one must evaluate it as fast as possible under some accuracy constraint.

When looking for good approximation polynomials, "good" has various possible meanings. For arbitrary precision implementations, polynomials must be built at runtime, so "good" means "simple" (for both the polynomial and the error term). Typical techniques in this case are based on Taylor or Chebyshev formulae. For fixed-precision implementations (for instance for the functions of the standard floating-point mathematical library), the polynomial is static, and we may afford to spend much more effort to build it. In this case, we may aim for better polynomials, in the sense that they minimize the approximation error over a complete interval: such polynomials are given by Remez' algorithm [56]. However, the coefficients of Remez polynomials will be arbitrary reals, and for implementation purpose we are more interested in the class of polynomials with machine-representable coefficients. An even better polynomial is therefore one that minimizes the approximation error among this class, a problem addressed in the Sollya toolbox developed in Arénaire (http://sollya.gforge.inria.fr/). In some cases it is useful to impose even more constraints on the polynomial. For instance, if the function is even, one classically wants to force to zero the coefficients of the odd powers in its polynomial approximation. Although this may require a higher degree approximation for the same accuracy, it reduces operation counts, and also increases the numerical stability of the evaluation.

Then, there are many ways to evaluate a polynomial, corresponding to many ways to rewrite it. The Horner scheme minimizes operation count and, in most practical cases, rounding errors, but it is a sequential scheme entailing a long execution time on modern processors. There exists parallel evaluation schemes that improve this latency, but degrade operation count and accuracy. The optimal scheme depends on details of the target architecture, and is best found by programmed exploration, as demonstrated by Intel on Itanium, and by Arénaire on the ST200 processor.

Thus, both polynomial approximation and polynomial evaluation illustrate the need for "meta-algorithms": i.e., algorithms designed to build arithmetic algorithms. In our example, the meta-algorithms in turn rely on linear algebra, integer linear programming, and Euclidean lattices. Other approaches may also lead to successful meta-algorithms, for instance the SPIRAL project (http://www.spiral.net/) uses algebraic rewriting to implement and optimize linear transforms. This approach has potential in arithmetic design, too.

3.5. Euclidean Lattice Reduction and Applications

A Euclidean lattice is the set of *integer* linear combinations of a finite set of real vectors. Typically, lattices occur when linear algebra questions are asked with discreteness constraints. In the last decade, they have become a classical ingredient in the computer arithmetic toolbox, along with other number-theoretic techniques (continued fractions, diophantine approximation, etc.). Indeed, integers (scaled by powers of the radix) are the essence of the fixed-point and floating-point representations of the real numbers. If the macroscopic properties of floating-point numbers are close to those of the real numbers, the finer properties are definitely related to questions over the integers. Thus, lattices have been successfully used in computer arithmetic to find constrained polynomial approximations to functions, and to attack the Table Maker's Dilemma. They have a potential for further arithmetic applications, for instance the design of digital filters.

Besides, the algorithms on Euclidean lattices are a rich experimentation laboratory for different types of arithmetics. The basis vectors are often represented exactly with long integer arithmetic. Furthermore, the fastest algorithms find the operations to be performed on the basis vectors via approximate computations, typically an approximate Gram-Schmidt orthogonalisation. These approximate computations may be performed with fixed-precision or arbitrary precision floating-point arithmetics. In some time-consuming applications of lattice algorithms, such as cryptanalyses of variants of RSA or lattice-based cryptosystems, integer linear programming, or even for solving the Table Maker's Dilemma, the practical run-time is of utmost importance. This motivates strong optimizations for the underlying arithmetics.

Further, aside from this strong relationship between lattices and arithmetics, the understanding of lattice-based cryptology is developing at a quick pace; making it efficient while remaining secure will require a thorough study, which must involve experts in both arithmetics and cryptology.

3.6. Reliability and Accuracy

Having basic arithmetic operators that are well-specified by standards leads to two directions. The first is to provide a guarantee that the implementations of these operators match their specification. The second is to use these operators as building blocks of well-specified computations, in other words to build upon these operators to obtain guarantees on the results of larger computing cores.

The approaches used to get such a guarantee vary greatly. Some computations are performed exactly, and in this case the results are considered to be intrinsically correct. However, exact values may not be finitely representable in the chosen number system and format: they must then be approximated. When an approximate value is computed using floating-point arithmetic, the specification of this arithmetic is employed to establish a bound on the roundoff errors, or to check that no exceptional situation occurred. For instance, the IEEE-754 standard for floating-point arithmetic implies useful properties, e.g., Dekker's error-free multiplication for various radices and precisions, the faithfulness of Horner's polynomial evaluation, etc.

Another possibility is that a simple final computation, still performed using floating-point arithmetic, enables to check whether a computed result is a reasonable approximation of the exact (unknown) result. Typically, to check that, for instance, a computed matrix R is close to the inverse of the initial matrix A, it suffices to check whether the product RA is close enough to the identity matrix. Such a simple, a posteriori, computation is called a *certificate*.

When considering more complicated functions, e.g., elementary functions, another issue arises. These functions have to be approximated, in general by polynomials. It no longer suffices to bound the rounding errors of the computations and check that no underflow/overflow may occur. One also has to take into account the approximation errors: certifying tight error bounds is quite a challenge. One usually talks of *verified computations* in this case.

Safety is typically based on interval arithmetic: what is computed is an interval which provably encloses the sought values. Naive interval arithmetic evaluates an expression as it is written, which does not take into account the dependencies between variables. This leads to irrelevant interval bloat. To address this problem, a solution is sometimes to rewrite the expression, a technique used for instance by the Gappa tool (http://gappa.gforge.inria.fr/) initially developed in Arénaire. Another systematic method is to use extensions to interval arithmetic. For instance, affine arithmetic has been used to optimize the data-path width of FPGA computing cores, and is also used in the Fluctuat tool to diagnose numerical instabilities in programs. When working with functions, Taylor models are a relevant extension: they represent a function as the sum of a polynomial of fixed degree and of an interval enclosing all errors (approximation as well rounding errors). This approach is very useful for computations involving function approximations, and has for instance been used successfully for the computation of the supremum norm of a function in one variable. The issue here is to devise algorithms that do not overestimate too much the result. It may be necessary to mix interval arithmetic and variable precision to reach the required level of guarantee and accuracy. In general, determining the right precision is difficult: the precision must be high enough to yield accurate results, but not too high since the computing time increases with the computing precision.

The complexity of some computer arithmetic algorithms, the intrinsic complexity of the floating-point model, the use of floating-point for critical applications, strongly advocate for the use of *formal proof* in computer arithmetic: a proof checker checks every step of the proof obtained by any means mentioned above. Even circuit manufacturers often provide a formal proof of the critical parts of their floating-point algorithms. For instance, the Intel divide and square root algorithms for the Itanium were formally proven. The expertise of the French community (which includes several ex-Arénaire members) in proving floating-point algorithms is well recognized. However, even the lower properties of the arithmetic are still challenging. For instance, with the specification of decimal arithmetic in the new version of the IEEE 754 standard, many theorems established in radix two have to be generalized to other radices.

ATEAMS Project-Team

3. Scientific Foundations

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 and to transform it. Each analysis or transformation begins with fact extraction. After the we may analyze specific software systems or large bodies of software systems. Our goal is to improve software systems by learning to understand the causes of complexity.

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 want to explore is the use of syntax-related program patterns for fact extraction. An early proposal for such a pattern-based approach is described in: a fixed base language (either C or PL/1 variant) is extended 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 closures) and the performance problems encountered, this approach has not seen wider use.

Another approach is proposed by de Moor and colleagues and uses path expressions on the syntax tree to extract program facts and formulate queries on them. This approach builds on the work of Paige and attempts to solve a classic problem: how to incrementally update extracted program facts (relations) after the application of a program transformation.

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 real (legacy) programming languages, which is highly relevant for experimental research and validation.

3.2.1. Goals

The main goal is to replace labour-intensive manual programming of fact extractors by automatic generation from annotated grammars or other concise and formal notation. 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. 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. Relational paradigm

For any source code analysis or transformation, after fact extraction comes elaboration, aggregation or other further analyses of these facts. For fact analysis, we base our entire research on the simple formal concept of a “relation”.

There are at least three lines of research that have explored the use of relations. First, in SQL, n-ary relations are used as basic data type and queries can be formulated to operate on them. SQL is widely used in database applications and a vast literature on query optimization is available. There are, however, some problems with SQL in the applications we envisage: (a) Representing facts about programs requires storing program fragments (e.g., tree-structured data) and that is not easy given the limited built-in datatypes of SQL; (b) SQL does not provide transitive closures, which are essential for computing many forms of derived information; (c) More generally, SQL does not provide fixed-point computations that help to solve sets of equations. Second, in Prolog, Horn clauses can be used to represent relational facts and inference rules for deriving new facts. Although the basic paradigm of Prolog is purely declarative, actual Prolog implementations add imperative features that increase the efficiency of Prolog programs but hide the declarative nature of the language. Extensions of Prolog with recursion have resulted in Datalog in many variations [AHV95]. In F(p)–L a Prolog database and a special-purpose language are used to represent and query program facts.

Third, in SETL, the basic data type was the set. Since relations can easily be represented as sets of tuples, relation-based computations can, in principle, be expressed in SETL. However, SETL as a language was very complicated and has not survived. However, work on programming with sets, bags and lists has continued well into the 90’s and has found a renewed interested with the revival of Lisp dialects in 2008 and 2009.

We have already mentioned the relational program representation by Linton. In Rigi, a tuple format (RSF) is introduced to represent untyped relations and a language (RCL) to manipulate them. Relational algebra is used in GROK, Crocopat and Relation Partition Algebra (RPA) to represent basic facts about software systems and to query them. In GUPRO graphs are used to represent programs and to query them. Relations have also been proposed for software manufacture, software knowledge management, and program slicing. Sometimes, set constraints are used for program analysis and type inference. More recently, we have carried out promising experiments in which the relational approach is applied to problems in software analysis and feature analysis. Typed relations can be used to decouple extraction, analysis and visualization of source code artifacts. These experiments confirm the relevance and viability of the relational approach to software analysis, and also indicate a certain urgency of the research direction of this team.

3.3.1. Goals

- New ideas and techniques for the efficient implementation of a relation-based specification formalism.
- Design and prototype implementation of a relation-based specification language that supports the use of extracted facts (Rascal).
- We target at uniform reformulations of existing techniques and algorithms for software analysis as well as the development of new techniques using the relational paradigm.
- We apply the above in the reformulation of refactorings for Java and domain specific languages.

3.4. Refactoring and Transformation

The final 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.

3.4.1. 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.

3.5. 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.

3.5.1. 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.

CAIRN Project-Team

3. Scientific Foundations

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 [1].

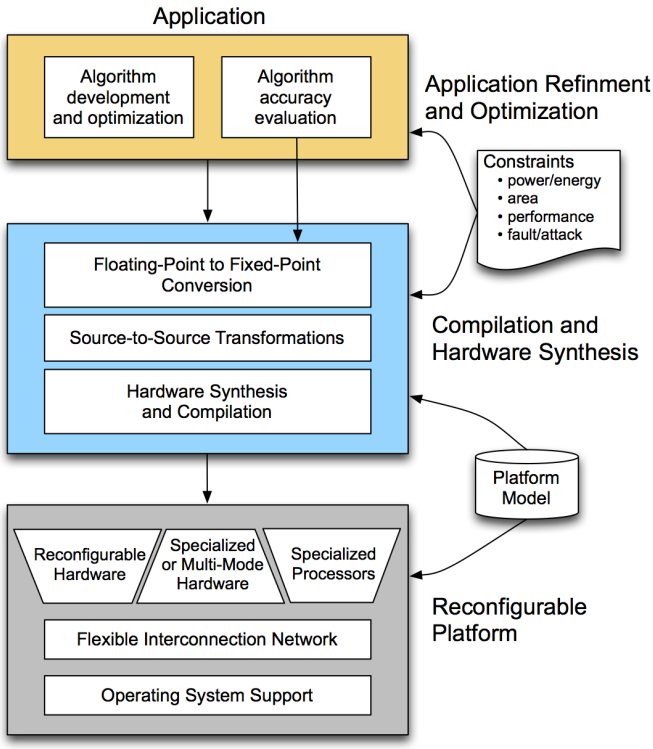


Figure 1. CAIRN*'s general design flow and related research themes*

Figure 1 shows the global design flow that 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).

[1] 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 [120], [104], [85], [125]). 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 [100][12], CRIP [88], Adres [112] or others [122]. These works have led to commercial products such as the Extreme Processor Platform (XPP) [89] from PACT or Montium [2] 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 [103], or of *field-configurable elements* such as the Xirisc processor [110] 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) [111]: communication, memory management, task scheduling [99], [92][1] and task placement [19]. 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 [118]. 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 CAIRN we 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.

[2] http://www.recoresystems.com/technology/montium-technology

- 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 [115], [91], [121], [124]. 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 [106], [115], [114]. 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 processor by extending their instruction set to speed-up compute 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 [93]. 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 [108], [113] 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 [107], [96]. 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 [97], [119]. 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 [101], [109] and have shown their efficiency in terms of energy consumption [98]. 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 [87]. 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 [94], 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. Scientific Foundations

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 [43]. 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).

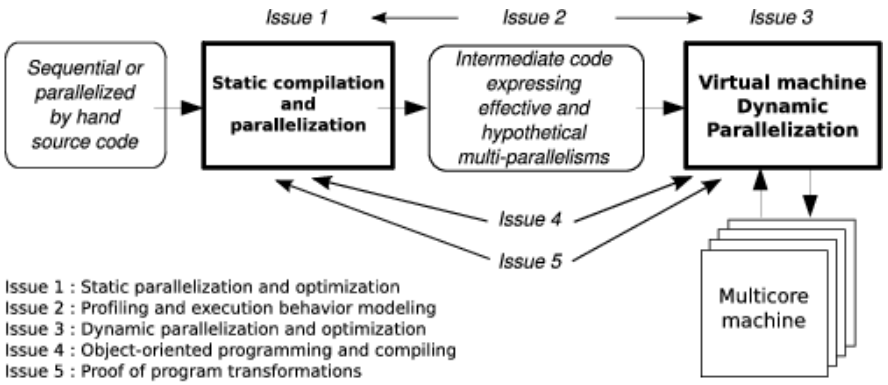


Figure 1. Automatic parallelizing steps for multicore architectures

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, Alexandra Jimborean.

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 [26]. 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. They are described below.

3.2.1. State of the art

Upstream, an easy interprocedural dependence analysis allows to handle complete programs (but recursivity: recursive functions must be transformed into iterative functions). Concerning iterative control we will use the polyhedral model, a formalism developped these last two decades, which allows to represent the execution of a loop nest by scanning a polytope.

When compiling an application, if it contains loop nests with affine bounds accessing scalars or arrays accessed using affine functions, the polyhedral model allows to:

- compute the dependence graph, which describes the order in which the dependent instructions must be executed [34];
- generate a schedule, which extracts some parallelism from the dependence graph [35], [36];
- generate an allocation, which assigns a processor (or a core) to a set of iterations of the loop nest to be scanned.

This last allocation step needs a thorough knowledge of the target architecture, as many crucial choices will result in performance hazards: for example, the volume and flow of inter-processor communications and synchronization; the data locality and the effects of the TLB (Translation Lookaside Buffer) and the various cache levels and distributions; or the register allocation optimizations. There are many techniques to control these parameters, and each architecture needs specific choices, of a valid schedule, of a parallel loop iterations distribution (bloc-, cyclic-, or tiled), of a loop-unrolling factor, as well as a memory data layout and a prefetch strategy (when available). They require powerful mathematical tools, such as counting the number of integer points contained in a parametric polytope.

Our own contributions in this area are significant. Concerning schedule and data placement, we proposed new advances in minimizing the number of communications for parallel architectures [54] and in cache access optimizations [53] [8]. We also proposed essential advances in parametric polytope manipulation [9], [5], developped the first algorithm to count integer points in a parametric polytope as an Ehrhart polynomial [3], and proposed successive improvements of this algorithm [10] [65]. We implemented these results in the free software *PolyLib*, utilized by many researchers around the world.

3.2.2. Adapting parallelization to multicore architecture

The first research direction to be explored is multicore specific efficient optimizations. Indeed, multicore architectures need specific optimizations, or we will get underlinear accelerations, or even decelerations. Multicore architectures may have the following properties: specific memory hierarchy, with distributed low-level cache and (possibly semi-) shared high level caches; software-controlled memory hierarchies (memory hints, local stores or scratchpads for example); optimized access to contiguous memory addresses or to separate memory banks; SIMD or vectorial execution in groups of cores, and synchronous execution; higher register allocation pressure when several threads use the same hardware (as in GPGPUs for example); etc.

A schedule and an allocation must be chosen wisely in order to obtain good performances. On NVIDIA GPGPUs, using the CUDA language, Baskaran et al. [25] obtained interesting results that have been implemented in their PLuTo compiler framework. However, they are based on many empirical and imprecise techniques, and require simulations to fine-tune the optimizations: they can be improved. Memory hierarchy efficient control is a cornerstone of tomorrow's multicore architectures performance. Compiler-optimizers have to evolve to meet this requirement.

Simulation and (partial-) profiling may however remain necessary in some cases, when static analysis reaches its intrinsic limits: when the execution of a program depends on dynamic parameters, when it uses complex pointer arithmetic, or when it performs indirect array accesses for example (as is often the case in *while* loops, out of the scope of the classical polyhedral model). In these cases, the compiler should rely on the profiler, and generate a code that interacts with the dynamic optimizer. This is the link with issues 2 and 3 of this research project.

3.2.3. Expressing many potential parallelisms

The dynamic optimizer (issue 3) must be able to exploit various parallel codes to compare them and the best one to choose, possibly swapping from a code to another during execution. The compiler must therefore generate different potentially efficient versions of a code, depending on fixed parameters such as the schedule or the data layout, and dynamic parameters such as the tile size or the unrolling factor.

The compiler then generates many variants of *effective* parallelism, formally proved by the static analyzer. It may also generate variants of code that have not been formally validated, due to the analyzer limits, and that have to be checked during execution by the dynamic optimizer: *hypothetical* parallelism. Hypothetical parallelism could be expressed as a piece of code, valid under certain conditions. Effective and hypothetical parallelisms are called *potential parallelism*. The variants of potential parallelism will be expressed in an intermediate language that has to be discovered.

Using compiler directives is an interesting way to define this intermediate language. Among the usual directives, we distinguish schedule directives for shared memory architectures (such as the OpenMP [1] *parallel* directive), and placement directives for distributed memory architectures (for example the HPF[2] *ALIGN* directive). These two types of directives are conjointly necessary to take full profit of multicore architectures. However, we have to study their complementarity and solve the interdependence or conflict that may arise between them. Moreover, new directives should allow to control data transfers between different levels of the memory hierarchy.

[1] http://www.openmp.org
[2] http://hpff.rice.edu

We are convinced that the definition of such a language is required in the next advances in compilation for multicore architectures, and there does not exist such an ambitious project to our knowledge. The OpenCL project [3], presented as an general-purpose and efficient multicore programming environment, is too low-level to be exploitable. We propose to define a new high level language based on compilation directives, that could be used by the skilled programmer or automatically generated by a compiler-optimizer (like OpenMP, recently integrated in the *gcc* compiler suite).

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.

3.3.1. Selective profiling and interaction with the compiler

In its simplest form, studying a given program's run time behavior consists in collecting and aggregating statistics, *e.g.*, counting how many times routines or basic blocks are executed, or counting the number of cache misses during a certain portion of the execution. In some cases, data can be collected about more abstract events, like the garbage-collector frequency or the number and sizes of sent and received messages. Such measures are relatively easy to obtain, are frequently used to quantify the benefits of some optimization, and may suggest some way to improve performance. These techniques are now well-known, but mostly for sequential programs.

These global studies have often been complemented by local, targeted techniques focused on some program portions, *e.g.*, where static techniques remain inconclusive for some fixed duration. These usages of profiling are usually strongly related to the optimization they complement, and are set up either by the compiler or by the execution environment. Their results may be used immediately at run time, in which case they are considered a form of run time optimization [1]. They can also be used offline to provide hints to a subsequent compilation cycle, in which case they constitute a form of profile-guided compilation, a strategy that is common in general purpose compilers.

For instance, in the context where a set of possible parallelizations have been provided by the compiler (see issue 1), a profiling component can easily be made responsible for testing some relevant condition at run time (*e.g.*, that depends on input data) and for selecting the best between various versions of the code. Beyond such simple tasks, we expect that profiling will, at the beginning of the execution, have enough resources to conduct more elaborate analyzes. We believe that combining an "open" static analysis with an integrated profiling component is a promising approach, first because it may relieve the programmer of a large part of the tedious task of implementing the distribution of computations, and second to free the compiler of the obligation to choose between several optimizations in the absence of enough relevant data. The main open question here is to define precisely the respective roles of the compiler and the profiler, and also the amount and nature of information the former can transmit to the latter.

[3] http://www.khronos.org/opencl

3.3.2. Profiling and dynamic optimization

In the context of dynamic optimization, that is, when the compiler's abilities have been exhausted, a profiler can still do useful work, provided some additional capabilities [1]. If it is able to instrument the code the way, *e.g.*, a PIN-tool does [55], it has access to the whole program, including libraries (or, for example, the code of a low-level library called from a scripting language). This means that it has access to portions of the program that were not under the compiler's control. The profiler can then perform dynamic inter-procedural analyzes, for instance to compute dependencies to detect parallelism that wasn't apparent at compile time because of a function call in the body of a loop. More generally, if the profiler is able to reconstruct at run time some representation of the whole program, as in [74] for example, it is possible to let it search for any construct that can be optimized and/or parallelized in the context of the current execution. Several virtual machines, *e.g.*, for Java or Microsoft CLR, have opened this way of optimizing programs, probably because virtual machines need to maintain an intermediate, structured representation of the running program.

The possibility of running programs on architectures that include a large number of computing cores has given rise to new abstractions [72], [46], [29]. Transactional memories, for instance, aim at simplifying the management of conflicting concurrent accesses to a shared memory, a notoriously difficult problem [48]. However, the performance of a transaction-based application heavily depends on its dynamic behavior, and too many conflicting accesses and rollbacks, severely affect performance. We bet that the need for multicore specific programming tools will lead to other abstractions based on speculative execution. Because of the very nature of speculation, all these abstractions will require run time evaluation, and maybe correction, to avoid pathological cases. The profiler has a central role here, because it can be made responsible for diagnosing inefficient use of speculative execution, and for taking corrective action, which means that it has to be integrated to the execution environment. We also think that the large scope and almost infinite potential uses of a profiling component may well suggest new parallel program abstractions, specially targeted at run time evaluation and adaptation.

3.3.3. Run time program modeling

When profiling goes beyond simple aggregation of counts, it can, for example, sample a program's behavior and split its execution into phases. These phases may help target a subsequent evaluation on a new architecture [66]. When profiling instruments the whole program to obtain a trace, *e.g.*, of memory accesses, it is possible to use this trace for:

- simulation, *e.g.*, by varying the parameters of the memory hierarchy,
- for modeling, *e.g.*, to reconstruct some specific model of the program [74], or to extract dynamic dependencies that help identifying parallel sections [62].

Handling such large execution traces, and especially compressing them, is a research topic by itself [30], [57]. Our contribution to this topic [7] is unusual in that the result of compression is a sequence of loop nests where memory accesses and loop bounds are affine functions of the enclosing loop indices. Modeling a trace this way leads to slightly better average compression rates compared to other, less expressive techniques. But more importantly, it has the advantage to provide a result in symbolic form, and this result can be further analyzed with techniques usually restricted to the static analysis of source code. We plan to apply, in the short term, similar techniques to the modeling of dynamic dependencies, so as to be able to automatically extract parallelism from program traces.

This kind of analysis is representative of a new kind of tools than could be named "parallelization assistants" [52], [62]. Properties that can't be detected by the compiler but that appear to hold in one or several executions of a program can be submitted to the programmer, maybe along a suitable reformulation of its program using some class of abstraction, *e.g.*, compiler directives. The goal is to provide help and guidance in adapting source code, in the same way a classical profiling tool helps pinpoint performance bottlenecks. Control and data dependencies are fundamental to such a tool. An execution trace provides an observed reality; for example a trace of memory addresses. If the observed dynamic dependencies provide a set of constraints, they also suggest a complete family of potential correct executions, be they parallel or sequential, and all these executions are equivalent to the reference execution. Being able to handle large traces, and representing them

in some manageable way, means being able to highlight medium to large grain parallelism, which is especially interesting on multicore architectures and often difficult for compilers to discover, for example because of the use of pointers and the difficulty of eliminating potential aliasing. This can be seen as a machine learning problem, where the goal is to recover a hidden structure from a large sequence of events. This general problem has various incarnations, depending on how much the learner knows about the original program, on the kind of data obtained by profiling, on the class of structures sought, and on the objectives of the analysis. We are convinced that such studies will enrich our understanding of the behavior of programs, and of the programming concepts that are really useful. It will also lead to useful tools, and will open up new directions for dynamic optimization.

3.4. Dynamic parallelization and optimization, virtual machine

Participants: Alexandra Jimborean, Philippe Clauss, Alain Ketterlin, Aravind Sukumaran-Rajam, Vincent Loechner.

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.

3.4.1. State of the art

Dynamic analysis and optimization, that is to say simultaneous to the program execution, have motivated a growing interest during the last decade, mainly because of the hardware architectures and applications growing complexity. Indeed, it has become more and more difficult to anticipate any program run simply from its source code, either because its control structures introduce some unknown objects before run (dynamic memory allocation, pointers, ...), or because the interaction between the target architecture and the program generates unpredictable behaviors. This is notably due to the appearance of more optimizing hardware units (prefetching units, speculative processing, code cache, branch prediction, etc.). With multicore architectures, this interest is growing even more. Works achieved in this area for mono-core processors have permitted to establish some classification of the so-called dynamic approaches, either based on the used methodologies or on the objectives.

The first objective for any dynamic approach is to extract some live information at runtime relying on a profiling process. This essential step is the main objective of issue 2 (see sub-section 3.3).

Identifying some “hotspots” thanks to profiling is then used for performance improvement optimizations. Two main approaches can be distinguished:

- the *profile-guided* approach, where analysis and optimization of profile information are performed off-line, that is to say statically. A first run is only performed to extract information for driving a re-compilation. Related to this approach, *iterative compilation* consists in running a code that has been transformed following different optimization possibilities (nature and sequencing of the applied optimizations), and then in re-compiling the transformed code guided by the collected performance information, and so on until obtaining a “best” program version. In order to promote a rapid convergence towards a better solution, some heuristics or some machine learning mechanisms are used [21], [61], [60]. The main drawback of such approaches relates to the quality of the generated code which depends on the reference profiled execution, and more precisely on the used input data set, but also on the used hardware.
- the *on-the-fly* approach consists in performing all steps at each run (profiling, analysis and transformation). The main constraint of this approach is that the time overhead has to be widely compensated

by the benefits it generates. Several works propose such approaches dedicated to specific optimizations. We personally successfully implemented a dynamic data prefetching system for the Itanium processor [1].

Although all these works provided some efficient dynamic mechanisms, their adaptation to multicore architectures yields difficult issues, and even challenges them. It is indeed necessary to control interactions between simultaneous tasks, imposing an additional complexity level which can be fateful for a dynamic system, while becoming too costly in time and space.

Some dynamic parallelizing techniques have been proposed in the last years. They are mainly focusing on parallelizing loop-nests, as programs generally spend most of their execution time in iterative structures.

The LRPD test [64] is certainly one of the foundation strategies. This method consists in speculatively parallelizing loops. Privatization and reduction transformations are applied to promote a successful application of the strategy. During execution, some tests are performed to verify the speculation validity. In case of invalid speculation, the targeted loop is re-executed sequentially. However, the application range is limited to loops accessing arrays; pointers cannot be handled. Moreover the method is not fully dynamic since an initial static analysis is needed.

In [33], Cintra and Llanos present a speculative parallel execution mechanism for loops, where iteration chunks are executed in sliding windows of n threads. The loops are not transformed and the sequential schedule remains as a reference to define a total order on the speculative threads. In order to verify whether some dependencies are violated during the program run, all data structures qualified as speculative, that is to say those being accessed in read-write mode by the threads, are duplicated for each thread and tagged following those states: *not accessed, modified, exposed loaded* or *exposed loaded and later modified*. For example, a *read-after-write* dependency has been violated if a thread owns a data tagged as *exposed loaded* or *exposed loaded and modified*, and if a predecessor thread, following the sequential total order, owns the same data but tagged as *modified* or *exposed loaded and modified*, while this data has not yet been committed in main memory. Such an approach can be memory-costly as each shared data structure is duplicated. It can be tricky to adjust verification frequencies to minimize time overhead. Some other methods based on the same principle of verifying speculation relatively to the sequential schedule have been proposed recently as in [68], where each iteration of a loop is decomposed into a prologue, a speculative body and an epilogue. The speculative bodies are performed in parallel and each body completion induces a verification. This approach seems to be only well suited for loops which bodies represent significant computation time.

Another recent work is the development of SPICE [63] which is a speculative parallelizing system where an entire first run of a loop is initially observed. This observation serves in determining the values reached by some variables during the run. During a next run of the loop, several speculative threads are launched. They consider as initial values of some variables the values that have been observed at the previous run. If a thread reaches the starting value of another thread, it stops. Thus each thread performs a different portion of the loop. But if the loop behavior changes and if another thread starting value is never reached, the run goes on sequentially until completion.

The main limits of these propositions are:

- they do not alter the initial sequential schedule since always contiguous instruction blocks are speculatively parallelized;
- their underlying parallelism is out of control: the characteristics of the generated parallel schedule are completely unknown since they randomly depend on the program instructions, their dependencies and the target machine. If bad performance is encountered, no other parallelization solution can be proposed. Moreover, the effective instruction schedule occurring at program run can significantly vary from one run to another, hence leading to a confusing performance inconsistency.

A strategy that would uniquely be based on a transactional memory mechanism, with rollbacks in the case of data races, yields a totally uncontrolable parallelism where performance can not be ensured and not even strongly expected.

While being based on efficient prediction mechanisms, a better control over parallelization will permit to provide solutions that are well suited to a varying execution context and to parallelize portions of code that can be parallelized only in some particular context. It is indeed crucial to maximize the potential parallelism of the applications to take advantage of the forthcoming processors comprising several tens of cores.

3.4.2. General objective: building a virtual machine

As it has already been mentioned, dynamic parallelization and optimization can take place inside a virtual machine. All the research objectives that are presented in the following are related to its construction.

Notice that the term of "virtual machine" is employed to group a set of dynamic analysis and optimization mechanisms taking as input a binary code, eventually enriched with specific instructions. We refer to a process virtual machine which main role is dynamic binary optimization from one instruction set to the same instruction set. The taxonomy given in [67] includes this kind of virtual machine.

Notice that this virtual machine can run in parallel on the processor cores during the four initial phases (see figure 2), but also simultaneously to the target application, either by sharing some cores with light processes, or by using cores that are useless for the target application. It will also support a transactional memory mechanism, if available. However the foreseen parallelizing strategies do not depend on such a mechanism since our speculative executions are supposed to be as reliable as possible thanks to efficient prediction models, and since they are supported by a specific and higher level rollback mechanism. Anyway if available, a transactional memory mechanism would allow to take advantage of "nearly perfect" prediction models.

The virtual machine takes as input an intermediate code expressing several kinds of parallelism on several code extracts. Those kinds of parallelism are either effective, that is to say that the corresponding parallel execution is obviously semantically correct, or hypothetical, that is to say that there is still some uncertainty on the parallelism correctness. In this case, this uncertainty will have to be resolved at run time. This intermediate "multi-parallel" code is generated by the static parallelization described subsection 3.2 . It also contains generic descriptions of parallelizing or optimizing transformations which parameters will have to be instanciated by the virtual machine, thanks to its knowledge about the target architecture and the program run-time behavior.

3.4.3. Adaptation of the intermediate code to the target architecture

The virtual machine first phase is to adapt this intermediate code to the target multicore architecture. It consists in answering the following questions:

- What is the suitable kind of parallelism?
- What is the suitable parallel task granularity?
- What is the suitable number of parallel tasks?
- Can we take advantage of a specialized instruction set for some operations?
- What are the parameter values for some parallelization or optimization?

The multi-parallel intermediate code exhibits different parameters allowing to adapt some parallelizing and optimizing transformations to the target architecture. For example, a loop unrolling will be parametrized by the number of iterations to be unrolled. This number will depend, for example, on the number of available registers and the size of the instruction cache. A parallelizing transformation will depend on several possible parallel instruction schedules. One or several schedules will be selected, for example, depending on the kind of memory hierarchy and the cache sharing among cores.

Concerning hypothetical parallelism, this first phase will reduce the number of these propositions to solutions that are well suited to the target architecture. This phase also instruments the intermediate code in order to install the dynamic mechanisms related to profiling and speculative parallel execution.

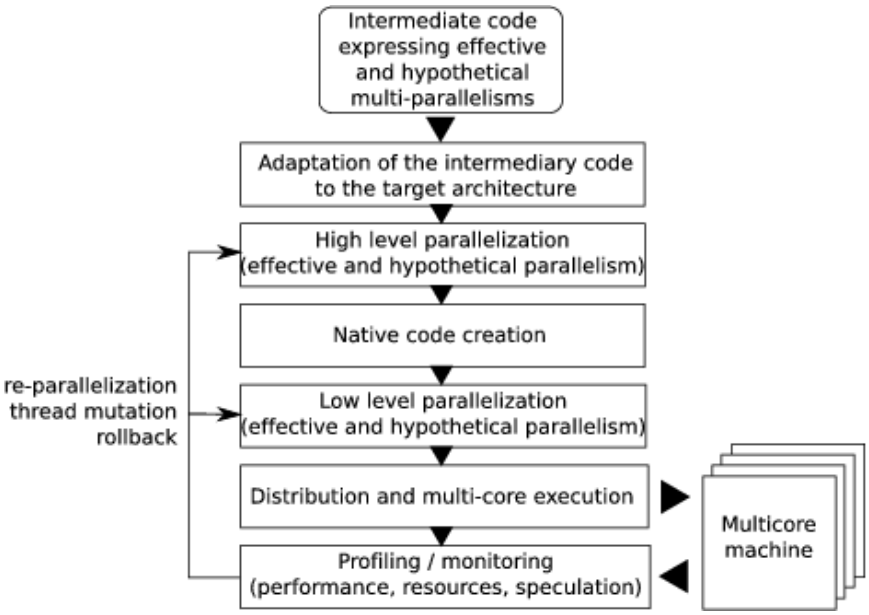


Figure 2. The virtual machine

3.4.4. High level parallelization and native code creation

From these target architecture related adaptations, a parallel intermediate code is generated. It contains instructions that are specific to the dynamic optimizing and parallelizing mechanisms, *i.e.*, instrumentation instructions to feed the profiling process as well as calls to speculative execution management procedures. A translation into native code executable by the target processor follows. This translation also allows to keep trace of the code extracts that have to be modified during the run.

3.4.5. Low level parallelization

The binary version of the code exhibits new parallelism and optimization sources that are specific to the instruction set and to the target architecture capabilities. Moreover, some dynamic optimizations are dedicated to specific instructions, or instruction blocks, as for example the memory reads which time performances can be dynamically improved by data prefetching [1]. Thus the binary code can be transformed and instrumented as well.

3.4.6. Distribution, execution and profiling

The so built executable code is then distributed among the processor cores to be run. During the run, the instrumentation instructions feed the profiler with information for execution monitoring and for behavior models construction (see subsection 3.3). An accurate knowledge of the binary code, thanks to the control of its generation, also permits at this step to dynamically control the insertion or deletion of some instrumentation instructions. Indeed it is important to manage execution monitoring through sampling based instrumentations in varying frequencies, following the changing behavior frequency (see in [1] and [73] a description of this kind of mechanism), as such instrumentations necessarily induce overheads that have to be minimized.

3.4.7. Re-parallelization, thread mutation or rollback

Depending on the information collected from instrumentation, and depending on the built prediction models, the profiling phase causes a re-transformation of some code parts, thus causing the mutation of the concerned

threads. Such re-transformation is done either on the binary code whether it consists in low level and small modifications, as for example the adjustement of a data prefetching distance, or on the intermediate code if it consists in a complete modification of the parallelizing strategy. For example, such a processing will follow the observation of a bad performance, or of a change in the computing resources availability, or will be caused by the completion of a dependency prediction model allowing the generation of a speculative parallelization. From such a speculative execution, a re-transformation can consist in rolling back to a sequential execution version when the considered hypothetical parallelism, and thus the associated prediction model, has been evaluated wrong.

3.5. Proof of program transformations for multicores

Participants: Éric Violard, Julien Narboux, Nicolas Magaud, Vincent Loechner, Alexandra Jimborean.

3.5.1. State of the art

3.5.1.1. Certification of low-level codes.

Among the languages allowing to exploit the power of multicore architectures, some of them supply the programmer a library of functions that corresponds more or less to the features of the target architecture : for example, CUDA [4] for the architectures of type GPGPU and more recently the standard OpenCL [5] that offers a unifying programming interface allowing the use of most of the existing multicore architectures or a use of heterogeneous aggregate of such architectures. The main advantage of OpenCL is that it allows the programmer to write a code that is portable on a large set of architectures (in the same spirit as the MPI library for multi-processor architectures). However, at this low level, the programming model is very close to the executing model, the control of parallelism is explicit. Proof of program correctness has to take into account low-level mechanisms such as hardware interruptions or thread preemption, which is difficult.

In [38], Feng *et al.* propose a logic inspired from the Hoare logic in order to certify such low-level programs with hardware interrupts and preempted threads. The authors specify this logic by using the meta-logic implemented in the Coq proof assistant [24].

3.5.1.2. Certification of a compiler.

The problem here is to prove that transformations or optimizations preserve the operational behaviour of the compiled programs.

Xavier Leroy in [27], [50] formalizes the analyses and optimizations performed by a C compiler: a big part of this compiler is written in the specification language of Coq and the executable (Caml) code of this compiler is obtained by automatic extraction from the specification.

Optimizing compilers are complex softwares, particularly in the case of multi-threaded programs. They apply some subtle code transformations. Therefore some errors in the compiler may occur and the compiler may produce incorrect executable codes. Work is to be done to remedy this problem. The technique of validation *a posteriori* [69], [70] is an interesting alternative to full verification of a compiler.

3.5.1.3. Semantics of directives.

As it was mentioned in subsection 3.2.3 , the use of directives is an interesting approach to adapt languages to multicore architectures. It is a syntactic means to tackle the increasing need of enriching the operational semantics of programs.

Ideally, these directives are only comments: they do not alter the correction of programs and they are a good means to improve their performance. They allow the separation of concerns: *correction* and *efficiency*.

However, using directives in that sense and in the context of automatic parallelization, raises some questions: for example, assuming that directives are not mandatory, how to ensure that directives are really taken into account? How to know if a directive is better than another? What is the impact of a directive on performance?

[4] http://www.nvidia.com/object/cuda_what_is.html
[5] http://www.khronos.org/opencl

In his thesis [40], that was supervised by Éric Violard, Philippe Gerner addresses similar questionings and states a formal framework in which the semantics of compilation directives can be defined. In this framework, any directive is encoded into one equation which is added to an algebraic specification. The semantics of the directives can be precisely defined via an order relation (called relation of *preference*) on the models of this specification.

3.5.1.4. Definition of a parallel programming model.

Classically, the good definition of a programming model is based on a semantic domain and on the definition of a "toy" language associated with a proof system, which allows to prove the correctness of the programs written in that language. Examples of such "toy" languages are CSP for control parallelism and $\mathcal{L}$ [28] for data parallelism. The proof systems associated with these two languages, are extensions of the Hoare logic.

We have done some significant works on the definition of data parallelism [11]. In particular, a crucial problem for the good definition of this programming model, is the semantics of the various syntactic constructs for data locality. We proposed a semantic domain which unifies two concepts: *alignment* (in a data-parallel language like HPF) and *shape* (in the data-parallel extensions of C).

We defined a "toy" language, called PEI, that is made of a small number of syntactic constructs. One of them, called *change of basis*, allows the programmer to exhibit parallelism in the same way as a placement or a scheduling directive [41].

3.5.1.5. Programming models for multicore architectures.

The multicore emergence questions the existing parallel programming models.

For example, with the programming model supported by OpenMP, it is difficult to master both correctness and efficiency of programs. Indeed, this model does not allow programmers to take optimal advantage of the memory hierarchy and some OpenMP directives may induce unpredictable performances or incorrect results.

Nowadays, some new programming models are experienced to help at designing both efficient and correct programs for multicores. Because memory is shared by the cores and its hierarchy has some distributed parts, some works aim at defining a hybrid model, between task parallelism and data parallelism. For example, languages like UPC (Unified Parallel C) [6] or Chapel [7] combine the advantages of several programming paradigms.

In particular, the model of memory transactions (or transactional memory [47]) retains much attention since it offers the programmer a simple operational semantics including a mutual exclusion mechanism which simplifies program design. However, much work remains to define the precise operational meaning of transactions and the interaction with the other languages features [56]. Moreover, this model leaves the compiler a lot of work to reach a safe and efficient execution on the target architecture. In particular, it is necessary to control the atomicity of transactions [39] and to prove that code transformations preserve the operational semantics.

3.5.1.6. Refinement of programs.

Refinement [22], [42] is a classical approach for gradually building correct programs: it consists in transforming an initial specification by successive steps, by verifying that each transformation preserves the correctness of the previous specification. Its basic principle is to derive simultaneously a program and its own proof. It defines a formal framework in which some rules and strategies can be elaborated to transform specifications written by using the same formalism. Such a set of rules is called a *refinement calculus*.

Unity [32] and Gamma [23] are classical examples of such formalisms, but they are not especially designed for refining programs for multicore architectures. Each of these formalisms is associated with a computing model and thus each specification can be viewed as a program. Starting with an initial specification, a proof logic allows a user to derive a specification which is more suited to the target architecture.

[6] http://upc.gwu.edu
[7] http://chapel.cs.washington.edu

Refinement applies for the programming of a large range of problems and architectures. It allows to pass the limitations of the polyhedral model and of automatic parallelization. We designed a refinement calculus to build data parallel programs [71].

3.5.2. Main objective: formal proof of analyses and transformations

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 [44]. Proving transformations involving program which are not data-race free will constitute a longer term research goal.

3.5.3. Proof of transformations in the polyhedral model

The main code transformations used in the compiler and the virtual machine are those carried out in the polyhedral model [49], [37]. We will use the Coq proof assistant to formalize proofs of analyses and transformations based on the polyhedral model. In [31], Cachera and Pichardie formalized nested loops in Coq and showed how to prove *properties* of those loops. Our aim is slightly different as we plan to prove *transformations* of nested loops in the polyhedral model. We will first prove the simplest unimodular transformations, and later we will focus on more complex transformations which are specific to multicore architectures. We will first study scheduling optimizations and then optimizations improving data locality.

3.5.4. Validation under hypothesis

In order to prove the correction of a code transformation T it is possible to:

- prove that T is correct in general, *i.e.*, prove that for all x, $T(x)$ is equivalent to x.
- prove *a posteriori* that the applied transformation has been correct in the particular case of a code c.

The second approach relies on the definition of a program called *validator* which verifies if two pieces of program are equivalent. This program can be modeled as a function V such that, given two programs c_1 and c_2, $V(c_1, c_2) = true$ only if c_1 has the same semantics as c_2. This approach has been used in the field of optimizations certification [59], [58]. If the validator itself contains a bug then the certification process is broken. But if the validator is proved formally (as it was achieved by Tristan and Leroy for the Compcert compiler [69], [70]) then we get a transformed program which can be trusted in the same way as if the transformation is proved formally.

This second approach can be used only for the *effective parallelism*, when the static analysis provides enough information to parallelize the code. For the *hypothetical parallelism*, the necessary hypotheses have to be verified at run time.

For instance, the absence of aliases in a piece of code is difficult to decide statically but can be more easily decided at run time.

In this framework, we plan to build a *validator under hypotheses*: a function V' such that, given two programs c_1 and c_2 and an hypothesis H, if $V'(c_1, c_2, H) = true$, then H implies that c_1 has the same semantics as c_2. The validity of the hypothesis H will be verified dynamically by the virtual machine. This verification process, which is part of the virtual machine, will have to be proved as correct as well.

3.5.5. Rejecting incorrect parallelizations

The goal of the project is to exhibit potential parallelism. The source code can contain many sub-routines which could be parallelized under some hypothesis that the static analysis fails to decide. For those optimizations, the virtual machine will have to verify the hypotheses dynamically. Dynamically dealing with the potential parallelism can be complex and costly (profiling, speculative execution with rollbacks). To reduce the overhead of the virtual machine, we will have to provide efficient methods to rule out quickly incorrect parallelism. In this context, we will provide hypotheses which are easy to check dynamically and which can tell when a transformation cannot be applied, *i.e.*, hypotheses which are sufficient conditions for the non-validity of an optimization.

CARAMEL Project-Team

3. Scientific Foundations

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-known algorithms for factoring integers or for computing discrete logarithms in finite fields have a subexponential complexity, whereas the best known attack 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 or the BSI, 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 (HFE, 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 [17], 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), and a large deployment is to be expected in the next years.

Despite the raise of elliptic curve cryptography and the variety of more or less mature other 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 want to 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.

We also plan to investigate and promote the use of pairing-based and genus-2 cryptosystems. For pairings, this is mostly a question of how efficient can such a system be in software, in hardware, and using all the tools from fast implementation to the search for adequate curves. For genus 2, as said earlier, constructing an efficient cryptosystem requires some more fundamental questions to be solved, namely the point-counting problem.

We summarize in the following table the aspects of public-key cryptography that we address in the CARAMEL team.

public-key primitive	cryptanalysis	design	implementation
RSA	X	–	–
Finite Field DLog	X	–	–
Elliptic Curve DLog	–	–	Soft
Genus 2 DLog	–	X	Soft
Pairings	X	X	Soft/Hard

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 have more and more 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 possible hardware platforms intended, 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 others goals, since they all rely more or less on integer or polynomial arithmetic.

CARTE Project-Team

3. Scientific Foundations

3.1. Computer Virology

From a historical point of view, the first official virus appeared in 1983 on Vax-PDP 11. At the very same time, a series of papers was published which always remains a reference in computer virology: Thompson [70], Cohen [39] and Adleman [29]. The literature which explains and discusses practical issues is quite extensive [43], [45]. 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 [58] 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, worms are an autonous self-replicating malware and so fall 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 [42], [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 [68]with continuous time.

3.3. Rewriting

The rewriting paradigm is now widely used for specifying, modelizing, programming and proving. It allows 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 [64].

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 [31], [41], [69].

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 [40]. 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 have proposed an approach using rewriting techniques to abstract program behaviors for detecting suspicious or malicious programs [32].

CASCADE Project-Team

3. Scientific Foundations

3.1. Provable Security

Since the beginning of public-key cryptography, with the seminal Diffie-Hellman paper [75], many suitable algorithmic problems for cryptography have been proposed and many cryptographic schemes have been designed, together with more or less heuristic proofs of their security relative to the intractability of the underlying problems. However, many of those schemes have thereafter been broken. The simple fact that a cryptographic algorithm withstood cryptanalytic attacks for several years has often been considered as a kind of validation procedure, but schemes may take a long time before being broken. An example is the Chor-Rivest cryptosystem [74], based on the knapsack problem, which took more than 10 years to be totally broken [90], whereas before this attack it was believed to be strongly secure. As a consequence, the lack of attacks at some time should never be considered as a full security validation of the proposal.

A completely different paradigm is provided by the concept of "provable" security. A significant line of research has tried to provide proofs in the framework of complexity theory (a.k.a. "reductionist" security proofs): the proofs provide reductions from a well-studied problem (factoring, RSA or the discrete logarithm) to an attack against a cryptographic protocol.

At the beginning, researchers just tried to define the security notions required by actual cryptographic schemes, and then to design protocols which could achieve these notions. The techniques were directly derived from complexity theory, providing polynomial reductions. However, their aim was essentially theoretical. They were indeed trying to minimize the required assumptions on the primitives (one-way functions or permutations, possibly trapdoor, etc) [78], without considering practicality. Therefore, they just needed to design a scheme with polynomial-time algorithms, and to exhibit polynomial reductions from the basic mathematical assumption on the hardness of the underlying problem into an attack of the security notion, in an asymptotic way. However, such a result has no practical impact on actual security. Indeed, even with a polynomial reduction, one may be able to break the cryptographic protocol within a few hours, whereas the reduction just leads to an algorithm against the underlying problem which requires many years. Therefore, those reductions only prove the security when very huge (and thus maybe unpractical) parameters are in use, under the assumption that no polynomial time algorithm exists to solve the underlying problem.

For a few years, more efficient reductions have been expected, under the denomination of either "exact security" [71] or "concrete security" [83], which provide more practical security results. The perfect situation is reached when one is able to prove that, from an attack, one can describe an algorithm against the underlying problem, with almost the same success probability within almost the same amount of time: "tight reductions". We have then achieved "practical security" [67].

Unfortunately, in many cases, even just provable security is at the cost of an important loss in terms of efficiency for the cryptographic protocol. Thus, some models have been proposed, trying to deal with the security of efficient schemes: some concrete objects are identified with ideal (or black-box) ones. For example, it is by now usual to identify hash functions with ideal random functions, in the so-called "random-oracle model", informally introduced by Fiat and Shamir [76], and later formalized by Bellare and Rogaway [70]. Similarly, block ciphers are identified with families of truly random permutations in the "ideal cipher model" [68]. A few years ago, another kind of idealization was introduced in cryptography, the black-box group, where the group operation, in any algebraic group, is defined by a black-box: a new element necessarily comes from the addition (or the subtraction) of two already known elements. It is by now called the "generic model" [82], [89]. Some works even require several ideal models together to provide some new validations [73].

More recently, the new trend is to get provable security, without such ideal assumptions (there are currently a long list of publications showing "without random oracles" in their title), but under new and possibly stronger computational assumptions. As a consequence, a cryptographer has to deal with the three following important steps:

computational assumptions, which are the foundations of the security. We thus need to have a strong evidence that the computational problems are reasonably hard to solve. We study several assumptions, by improving algorithms (attacks), and notably using lattice reductions. We furthermore contribute to the list of "potential" hard problems.

security model, which makes precise the security notions one wants to achieve, as well as the means the adversary may be given. We contribute to this point, in several ways:

- by providing a security model for many primitives and protocols, and namely group-oriented protocols, which involve many parties, but also many communications (group key exchange, group signatures, etc);
- by enhancing some classical security models;
- by considering new means for the adversary, such as side-channel information.

design of new schemes/protocols, or more efficient, with additional features, etc.

security proof, which consists in exhibiting a reduction.

For a long time, the security proofs by reduction used classical techniques from complexity theory, with a direct description of the reduction, and then a long and quite technical analysis for providing the probabilistic estimates. Such analysis is unfortunately error-prone. Victor Shoup proposed a nice way to organize the proofs, and eventually obtain the probabilities, using a sequence of games [88], [69], [84] which highlights the computational assumptions, and splits the analysis in small independent problems. We early adopted and developed this technique, and namely in [77]. We applied this methodology to various kinds of systems, in order to achieve the highest security properties: authenticity, integrity, confidentiality, privacy, anonymity. Nevertheless, efficiency was also a basic requirement.

However, such reductions are notoriously error-prone: errors have been found in many published protocols. Security errors can have serious consequences, such as loss of money in the case of electronic commerce. Moreover, security errors cannot be detected by testing, because they appear only in the presence of a malicious adversary.

Security protocols are therefore an important area for formal verification.

3.2. Cryptanalysis

Because there is no absolute proof of security, it is essential to study cryptanalysis, which is roughly speaking the science of code-breaking. As a result, key-sizes are usually selected based on the state-of-the-art in cryptanalysis. The previous section emphasized that public-key cryptography required hard computational problems: if there is no hard problem, there cannot be any public-key cryptography either. If any of the computational problems mentioned above turns out to be easy to solve, then the corresponding cryptosystems can be broken, as the public key would actually disclose the private key. This means that one obvious way to cryptanalyze is to solve the underlying algorithmic problems, such as integer factorization, discrete logarithm, lattice reduction, Gröbner bases, *etc.* Here, we mean a study of the computational problem in its full generality. The project-team has a strong expertise (both in design and analysis) on the best algorithms for lattice reduction, which are also very useful to attack classical schemes based on factorization or discrete logarithm.

Alternatively, one may try to exploit the special properties of the cryptographic instances of the computational problem. Even if the underlying general problem is NP-hard, its cryptographic instances may be much easier, because the cryptographic functionalities typically require a specific mathematical structure. In particular, this means that there might be an attack which can only be used to break the scheme, but not to solve the underlying problem in general. This happened many times in knapsack cryptography and multivariate cryptography. Interestingly, generic tools to solve the general problem perform sometimes even much better on cryptographic instances (this happened for Gröbner bases and lattice reduction).

However, if the underlying computational problem turns out to be really hard both in general and for instances of cryptographic interest, this will not necessarily imply that the cryptosystem is secure. First of all, it is not even clear what is meant exactly by the term *secure* or *insecure*. Should an encryption scheme which leaks the first bit of the plaintext be considered secure? Is the secret key really necessary to decrypt ciphertexts or to sign messages? If a cryptosystem is theoretically secure, could there be potential security flaws for its implementation? For instance, if some of the temporary variables (such as pseudo-random numbers) used during the cryptographic operations are partially leaked, could it have an impact on the security of the cryptosystem? This means that there is much more into cryptanalysis than just trying to solve the main algorithmic problems. In particular, cryptanalysts are interested in defining and studying realistic environments for attacks (adaptive chosen-ciphertext attacks, side-channel attacks, *etc.*), as well as goals of attacks (key recovery, partial information, existential forgery, distinguishability, *etc.*). As such, there are obvious connections with provable security. It is perhaps worth noting that cryptanalysis also proved to be a good incentive for the introduction of new techniques in cryptology. Indeed, several mathematical objects now considered invaluable in cryptographic design were first introduced in cryptology as cryptanalytic tools, including lattices and pairings. The project-team has a strong expertise in cryptanalysis: many schemes have been broken, and new techniques have been developed.

3.3. Symmetric Cryptography

Even if asymmetric cryptography has been a major breakthrough in cryptography, and a key element in its recent development, conventional cryptography (a.k.a. symmetric, or secret key cryptography) is still required in any application: asymmetric cryptography is much more powerful and convenient, since it allows signatures, key exchange, etc. However, it is not well-suited for high-rate communication links, such as video or audio streaming. Therefore, block-ciphers remain a fundamental primitive. However, since the AES Competition (which started in January 1997, and eventually selected the Rijndael algorithm in October 2000), this domain has become less active, even though some researchers are still trying to develop new attacks. On the opposite, because of the lack of widely admitted stream ciphers (able to encrypt high-speed streams of data), ECRYPT (the European Network of Excellence in Cryptology) launched the eSTREAM project, which investigated research on this topic, at the international level: many teams proposed candidates that have been analyzed by the entire cryptographic community. Similarly, in the last few years, hash functions [86], [85], [80], [81], [79], which are an essential primitive in many protocols, received a lot of attention: they were initially used for improving efficiency in signature schemes, hence the requirement of collision-resistance. But afterwards, hash functions have been used for many purposes, such as key derivation, random generation, and random functions (random oracles [70]). Recently, a bunch of attacks [72], [91], [92], [93], [94], [96], [95] have shown several drastic weaknesses on all known hash functions. Knowing more (how weak they are) about them, but also building new hash functions are major challenges. For the latter goal, the first task is to formally define a security model for hash functions, since no realistic formal model exists at the moment: in a way, we expect too much from hash functions, and it is therefore impossible to design such "ideal" functions. Because of the high priority of this goal (the design of a new hash function), the NIST has launched an international competition, called SHA-3 (similar to the AES competition 10 years ago), in order to select and standardize a hash function. Keccak has been officially chosen on October 2nd, 2012.

One way to design new hash functions may be a new mode of operation, which would involve a block cipher, iterated in a specific manner. This is already used to build stream ciphers and message authentication codes (symmetric authentication). Under some assumptions on the block cipher, it might be possible to apply the above methodology of provable security in order to prove the validity of the new design, according to a specific security model.

CASSIS Project-Team

3. Scientific Foundations

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 constraints [6] to 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. Scientific Foundations

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 [71].
- 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 [51], 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** [49], [74]. 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 [50]. 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 [3].

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) [42] 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 [72], [73] 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.,* [60], [59] have explored the algorithmic design space of contexts for benchmarks of significant size. Latter on, Milanova *et. al.,* [66] 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 [64] have extended the empiric evaluation of object-sensitivity using a BDD implementation allowing to cope with benchmarks otherwise out-of-scope. Besson and Jensen [46] 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 [65] and the race analysis of Naik, Aiken [67] and Whaley [68] 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 [55] 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 [63]. The proposed technique has been implemented in a tool called NIT [62]. Null pointer detection is also done by bug-detection tools such as FindBugs [62]. 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 [61], [41], linear

logic approaches to implicit computational complexity [43], cost model for Java byte code [37] based on size relation inference, and WCET computation by abstract interpretation based loop bound interval analysis techniques [52].

We have proposed an original approach for designing static analyses computing program costs: inspired from a probabilistic approach [75], 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 [48]. An example of cache miss analysis has been developed within this framework [79].

3.1.3. Semantic analysis for test case generation

The semantic analysis of programs can be combined with efficient constraint solving techniques in order to extract specific information about the program, *e.g.*, concerning the accessibility of program points and feasibility of execution paths [76], [54]. As such, it has an important use in the automatic generation of test data. Automatic test data generation received considerable attention these last years with the development of efficient and dedicated constraint solving procedures and compositional techniques [58].

We have made major contributions to the development of **constraint-based testing**, which is a two-stage process consisting of first generating a constraint-based model of the program's data flow and then, from the selection of a testing objective such as a statement to reach or a property to invalidate, to extract a constraint system to be solved. Using efficient constraint solving techniques allows to generate test data that satisfy the testing objective, although this generation might not always terminate. In a certain way, these constraint techniques can be seen as efficient decision procedures and so, they are competitive with the best software model checkers that are employed to generate test data.

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 [1] 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.

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], [45]. 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 [47]. **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) [69] 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" [70]. Many solutions have been proposed to reduce the size of the trusted computing base. In the *foundational proof carrying code* of Appel and Felty [40], [39], 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 [70] 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”

[1] 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 ?

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.* [38] 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 [36] 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 [5].

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 [56], [57]. We defined a simple and efficient algorithm [53] 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 [3]. This checker is now used to certify all analysis results produced by the regular completion tool as well as the optimised version of [44].

COMETE Project-Team

3. Scientific Foundations

3.1. Probability and information theory

Participants: Miguel Andrés, Nicolás Bordenabe, Konstantinos Chatzikokolakis, Ehab ElSalamouny, Sardaouna Hamadou, 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 be able to exploit such information.

The recent tendency is to use information theoretic approach to model the problem and define the leakage in a quantitative way. The idea is to consider that 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. Most of the proposals in the literature use *Shannon entropy*, which is the most established 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). We consider also other notions, in particular the Rényi min-entropy, which seem to be more appropriate for security in common scenarios like the one-try attacks.

3.2. The probabilistic asynchronous π-calculus

Participants: Konstantinos Chatzikokolakis, Marco Giunti, Catuscia Palamidessi, Frank Valencia, Lili Xu.

We will focus our efforts on a probabilistic variant of the asynchronous π-calculus, which is a formalism designed for mobile and distributed computation. A characteristic of our calculus is the presence of both probabilistic and nondeterministic aspects. This combination is essential to represent probabilistic algorithms and protocols, and express their properties in presence of unpredictable (nondeterministic) users and adversaries.

3.3. Expressiveness issues

Participants: Andrés Aristizábal, Catuscia Palamidessi, Luis Fernando Pino Duque, Frank Valencia.

We intend to study models and languages for concurrent, 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, taking also into account the issue of (efficient) implementability.

3.4. Concurrent constraint programming

Participants: Andrés Aristizábal, Sophia Knight, Luis Fernando Pino Duque, Frank Valencia.

Concurrent constraint programming (ccp) is a well-established process calculus [43] for modeling systems where agents interact by adding and asking information in a global store. 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 at the same time 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. 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. Enriching ccp with epistemic constructs, which will allow to reason about the knowledge of agents.

3.5. Model checking

Participants: Miguel Andrés, Catuscia Palamidessi.

We plan to develop model-checking techniques and tools for verifying properties of systems and protocols specified in the above formalisms.

Model checking addresses the problem of establishing whether the model (for instance, a finite-state machine) of a certain specification satisfies a certain logical formula.

We intend to concentrate our efforts on aspects that are fundamental for the verification of security protocols, and that are not properly considered in existing tools. Namely, we will focus on:

(a) the combination of probability and mobility, which is not provided by any of the current model checkers,

(b) the interplay between nondeterminism and probability, which in security present subtleties that cannot be handled with the traditional notion of scheduler,

(c) the development of a logic for expressing security (in particular privacy) properties.

Concerning the last point (the logic), we should capture both probabilistic and epistemological aspects, the latter being necessary for treating the knowledge of the adversary.

Logics of this kind have been already developed, but the investigation of the relation with the models coming from process calculi, and their utilization in model checking, is still in its infancy.

COMPSYS Project-Team

3. Scientific Foundations

3.1. Introduction

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 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. Such a combination 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 multi-cores 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 in this transformation chain, restricting to compilation (transforming a program to a program) for embedded processors 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. 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). In all aspects, the frontier between software and hardware is vanishing. For example, in terms of architecture, customized processors (with processor extension as 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. 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.

3.2. Back-End Code Optimizations for Embedded Processors

Participants: Quentin Colombet, Alain Darte, Fabrice Rastello.

Compilation is an old activity, in particular back-end code optimizations. We first give some elements that explain why the development of embedded systems makes compilation come back as a research topic. We then detail the code optimizations that we are interested in, both for aggressive and just-in-time compilation.

3.2.1. Embedded Systems and the Revival of Compilation & Code Optimizations

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 rediscovered the interest of 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), or (quasi) general-purpose processors (DSP for multimedia applications). Our cooperation with STMicroelectronics led us to investigate the last solution, as implemented in the ST100 (DSP processor) and the ST200 (VLIW DSP processor) family for example. Compilation and, in particular, back-end code optimizations find a second life in the context of such embedded computing systems.

At the heart of this progress is the concept of *virtualization*, which is the key for more portability, more simplicity, more reliability, and of course more security. This concept, implemented through binary translation, just-in-time compilation, etc., consists in hiding the architecture-dependent features as far as possible during the compilation process. It has been used for quite a long time for servers such as HotSpot, a bit more recently for workstations, and it is quite recent for embedded computing for reasons we now explain.

As previously mentioned, the definition of "embedded systems" is rather imprecise. However, one can at least agree on the following features:

- Even for processors that are programmable (as opposed to hardware accelerators), processors have some architectural specificities, and are very diverse;
- Many processors (but not all of them) have limited resources, in particular in terms of memory;
- For some processors, power consumption is an issue;
- In some cases, aggressive compilation (through cross-compilation) is possible, and even highly desirable for important functions.

This diversity is one of the reason why virtualization, which starts to be more mature, is becoming more and more common in programmable embedded systems, in particular through CIL (a standardization of MSIL). This implies a late compilation of programs, through just-in-time (JIT), including dynamic compilation. Some people even think that dynamic compilation, which can have more information because performed at run-time, can outperform the performances of "ahead-of-time" compilation.

Performing code generation (and some higher-level optimizations) in a late phase is potentially advantageous, as it can exploit architectural specificities and run-time program information such as constants and aliasing, but it is more constrained in terms of time and available resources. Indeed, the processor that performs the late compilation phase is, *a priori*, less powerful (in terms of memory for example) than a processor used for cross-compilation. The challenge is thus to spread the compilation process in time by deferring some optimizations ("deferred compilation") and by propagating some information for those whose computation is expensive ("split compilation"). Classically, a compiler has to deal with different intermediate representations (IR) where high-level information (i.e., more target-independent) co-exist with low-level information. The split compilation has to solve a similar problem where, this time, the compactness of the information representation, and thus its pertinence, is also an important criterion. Indeed, the 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 also a reason why static single assignment (SSA) is becoming specific to embedded compilation, even if it was first used for workstations. Indeed, SSA is a sparse (i.e., compact) representation of liveness information. In other words, 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.

3.2.2. Aggressive and Just-in-Time Optimizations of Assembly-Level Code

Compilation for embedded processors is difficult because the architecture and the operations are specially tailored to the task at hand, and because the amount of resources is strictly limited. For instance, 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, but also the special form of applications [20] generate many open problems. Our goal is to contribute to their understanding and their solutions.

As previously explained, compilation for embedded processors include both aggressive and just in time (JIT) optimizations. Aggressive compilation consists in allowing more time to implement costly solutions (so, looking for complete, even expensive, studies is mandatory): the compiled program is loaded in permanent memory (ROM, flash, etc.) and its compilation time is not significant; also, for embedded systems, code size and energy consumption usually have a critical impact on the cost and the quality of the final product. Hence, the application is cross-compiled, in other words, compiled on a powerful platform distinct from the target processor. Just-in-time compilation corresponds to compiling applets on demand on the target processor. For compatibility and compactness, the source languages are CIL or Java. The code can be uploaded or sold separately on a flash memory. Compilation is performed at load time and even dynamically during execution. Used heuristics, constrained by time and limited resources, are far from being aggressive. They must be fast but smart enough.

Our aim is, 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 for example in register allocation, in opcode selection, or in code placement for optimization of the instruction cache. Another example is the problem of removing the multiplexer functions (known as ϕ functions) that are inserted when converting into SSA form. These optimizations are usually done in the last phases of the compiler, using an assembly-level intermediate representation. In industrial compilers, they are handled in independent phases using heuristics, in order to limit the compilation time. Our initial goal was to develop a more global understanding of these optimization problems to derive both aggressive heuristics and JIT techniques, the main tool being the SSA representation.

In particular, we investigated the interaction of register allocation, coalescing, and spilling, with the different code representations, such as SSA. One of the challenging features of today's processors is predication [27], which interferes with all optimization phases, as the SSA form does. Many classical algorithms become inefficient for predicated code. This is especially surprising, since, beside giving a better trade-off between the number of conditional branches and the length of the critical path, converting control dependences into data dependences increases the size of basic blocks and hence creates new opportunities for local optimization

algorithms. One has to adapt classical algorithms to predicated code [29] and also to study the impact of predicated code on the whole compilation process.

As mentioned in Section 2.3 , a lot of progress has already been done in this direction in our past collaborations with STMicroelectronics. In particular, the goal of the Sceptre project was to revisit, in the light of SSA, some code optimizations in an aggressive context, i.e., by looking for the best performances without limiting, *a priori*, the compilation time and the memory usage. One of the major results of this collaboration was to propose to exploit SSA so as to design a register allocator in two phases, with one spilling phase relatively target-independent, then the allocator itself, which takes into account architectural constraints and optimizes other aspects (in particular, coalescing). This new way of considering register allocation has shown its interest for aggressive static compilation. But it offered three other perspectives:

- A simplification of the allocator, which again goes toward a more reliable compiler design, based on static single assignment.
- The possibility to handle the hardest part, the spilling phase, as a preliminary phase, thus a good candidate for split compilation.
- The possibility of a fast allocator, with a much higher quality than usual JIT approaches such as "linear scan", thus suitable for virtualization and JIT compilation.

These additional possibilities have been the heart of our research on back-end optimizations in Compsys II. The objective of the Mediacom project with STMicroelectronics was to address them. More generally, in Compsys II, our goal was to continue to develop our activity on code optimizations, exploiting SSA properties, following our two-phases strategy:

- First, revisit 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, exploit 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.

An important challenge was also to consider more code optimizations and more architectural features, such as registers with aliasing, predication, and, possibly in a longer term, vectorization/parallelization.

3.3. Program Analysis and Transformations for High-Level Synthesis

Participants: Christophe Alias, Alain Darte, Paul Feautrier, Laure Gonnord [Compsys/LIFL], Alexandru Plesco [Compsys/Zettice].

3.3.1. High-Level Synthesis Context

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 and time-to-market minimization are paramount. We believe that our expertise in compilation and automatic parallelization can 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 exist 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. 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. 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. It is 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.

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, are not directly applicable, firstly because they pay poor attention to locality, which is of paramount importance in hardware. Besides, their aim is to extract all the parallelism in the source code; they 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 both resource and locality 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.

In summary, as for our activity on back-end code optimizations, which is decomposed into two complementary activities, aggressive and just-in-time compilation, we focus our activity on high-level synthesis on two 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.
- Developing concepts and techniques in a more global view of high-level synthesis, starting from specification languages down to hardware implementation.

We now give more details on the program optimizations and transformations we want to consider and on our methodology.

3.3.2. Specifications, Transformations, Code Generation for High-Level Synthesis

Before contributing to high-level synthesis, 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 between FPGAs and the host processor.

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 flow from an input channel to the output. Here, data is 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. 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 (CRP, memory based). It is an interesting fact that several researchers have investigated translation from process networks [21] and to process networks [30], [31].

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 is best 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. In order to bypass this difficulty, one can 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. One can prove that the determinacy property still holds [22]. 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 [23]. Each statement (or group of statements) may be considered a process, and the source computation indicates where to implement communication channels. 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. Here, communications are associated to so-called uncut dependences, i.e., dependences which cross thread boundaries. In both approaches, the main question is whether the communications can be implemented as FIFOs, or need a reordering memory. One of our research directions will be to try to take advantage of the reordering allowed by dependences to force a FIFO implementation.

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 and for minimizing the length of the reuse vector [24], techniques for data allocation in scratch-pad memory, or techniques for folding multi-dimensional arrays [19]. 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? How does one take into account latency or throughput constraints, or bandwidth constraints for input and output channels? 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 tools) or as transformations to generate directly VHDL codes. 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.

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 naïve implementation may destroy the benefits of high-level optimization. There are two possibilities here as suggested before; one may target another high-level synthesis tool, or one may target directly VHDL. Each approach has its advantages and drawbacks. However, in both situations, all such tools require that the input program respects some strong constraints on the code shape, array accesses, memory accesses, communication protocols, etc. Furthermore, to get the tool to do what the user wants requires a lot of program tuning, i.e., of program rewriting. What can be automated in this rewriting process? Semi-automated?

CONTRAINTES Project-Team

3. Scientific Foundations

3.1. Rule-based Modeling Languages

Logic programming in a broad sense is a declarative programming paradigm based on mathematical logic with the following identifications:

$$program = logical\ formula,$$

$$execution = proof\ search,$$

In Constraint Satisfaction Problems (CSP), the logical formulae are conjunctions of constraints (i.e. relations on variables expressing partial information) and the satisfiability proofs are computed by constraint solving procedures.

In Constraint Logic Programming (CLP), the logical formulae are Horn clauses with constraints (i.e. one headed rules for the inductive definitions of relations on variables) and the satisfiability proofs combine constraint solving and clause resolution. Gnu-Prolog and its modular extension EMoP that we develop, belong to this family of languages. We use them for solving combinatorial problems and for implementing Biocham.

In Concurrent Constraint Programming (CCP), CLP resolution is extended with a synchronization mechanism based on constraint entailment. The variables play the role of transmissible dynamically created communication channels. An agent may add constraints to the store or read the store to decide whether a constraint guard is entailed by the current store. Sicstus-Prolog and SWI-Prolog belong the this family of languages. We use them for solving combinatorial optimization problems and defining new global constraints.

CCP execution can be identified to deduction in J.Y. Girard's Linear Logic by interpreting multisets of constraints and agents as tensor product conjunctions and guards and rules as linear implications [1]. The logical completeness of CCP in LL continues to hold when considering linear logic constraint systems, i.e. constraint systems where constraints can be consumed by implication. This extension, named Linear Logic Concurrent Constraint Programming (LLCC), allows for a non-monotonic evolution of the store of constraints and can encode multi-headed rules like the Constraint Handling Rules (CHR) language of T. Frühwirth.

All these rule-based languages, of increasing expressivity, involve some form of *multiset rewriting*. We have designed and continue developing the following modeling languages:

- Rules2CP, a rule-based modeling language for solving constraint optimization problems, developed for non-programmers,
- SiLCC, our experimental implementation of LLCC,
- the Biochemical Abstract Machine BIOCHAM, a rule-based modeling language dedicated to Systems Biology, in which biochemical reactions between multisets of reactants and products are expressed with multi-headed rules (somewhat similar to CHR rules) and augmented with *kinetic expressions* from which one can derive quantitative interpretations by Ordinary Differential Equations (ODE), Continuous-Time Markov Chains (CTMC) or Hybrid Automata.

3.2. Constraint Solving Techniques

Constraint propagation algorithms use constraints actively during search for filtering the domains of variables and reducing the search space. These domain reductions are the only way constraints communicate between each other. Our research involves different constraint domains, namely:

- booleans: binary decision diagrams and SAT solvers;

[1] F. Fages, P. Ruet, S. Soliman. *Linear concurrent constraint programming: operational and phase semantics*, in "Information and Control", 2001, vol. 165(1), pp.14-41.

- finite domains (bounded natural numbers): membership, arithmetic, reified [20], higher order and global constraints;
- reals: polyhedral libraries for linear constraints and interval methods;
- terms: subtyping constraints;
- graphs: subgraph epimorphism (SEPI) and isomorphism constraints; acyclicity constraint;
- Petri nets: P/T-invariants [5], siphons and traps [10];
- Kripke structures: temporal logic constraints (first-order Computation Tree Logic constraints over the reals).

We develop new constraints and domain filtering algorithms by using already existing constraint solving algorithms and implementations. For instance, we use the Parma Polyhedra Library PPL with its interface with Prolog for solving temporal logic constraints over the reals. Similarly, we use standard finite domain constraints for developing solvers for the new SEPI graph constraint.

3.3. Formal Methods for Systems Biology

At the end of the 90s, research in Bioinformatics evolved, passing from the analysis of the genomic sequence to the analysis of post-genomic interaction networks (expression of RNA and proteins, protein-protein interactions, transport, etc.). Systems biology is the name given to a pluridisciplinary research field involving biology, computer science, mathematics, physics, to illustrate this change of focus towards system-level understanding of high-level functions of living organisms from their biochemical bases at the molecular level.

Our group was among the first ones in 2002 to apply formal methods from computer science to systems biology in order to reason on large molecular interaction networks and get over complexity walls. The *logical paradigm for systems biology* that we develop can be summarized by the following identifications :

$$\textit{biological model} = \textit{rule-based transition system},$$

$$\textit{biological property} = \textit{temporal logic formula},$$

$$\textit{model validation} = \textit{model-checking},$$

$$\textit{model inference} = \textit{constraint solving}.$$

Rule-based dynamical models of biochemical reaction networks are composed of a reaction graph (bipartite graph with vertices for species and reactions) where the reaction vertices are given with kinetic expressions (mass action law, Michaelis-Menten, Hill, etc.). Most of our work consists in analysing the *interplay between the structure* (reaction graphs) *and the dynamics* (ODE, CTMC or hybrid interpretations derived from the kinetic expressions).

Besides this logical paradigm, we use the theory of abstract interpretation to relate the different interpretations of rule-based models and organize them in a hierarchy of semantics from the most concrete (CTMC stochastic semantics) to the most abstract (asynchronous Boolean transition system). This allows us to prove for instance that if a behavior is not possible in the Boolean semantics of the rules then it is not possible in the stochastic semantics for any kinetic expressions and parameter values. We also use the framework of abstract interpretation to formally relate rule-based reaction models to other knowledge representation formalisms such as, for instance, ontologies of protein functions, or influence graphs between molecular species. These formal methods are used to build models of biological processes, fit models to experimental data, make predictions, and design new biological experiments.

3.4. Tight Integration of In Silico and In Vivo Approaches

Bridging the gap between the complexity of biological systems and our capacity to model and predict systems behaviors is a central challenge in quantitative systems biology. We investigate using wet and dry experiments a few challenging biological questions that necessitate a tight integration between *in vivo* and *in silico* work. Key to the success of this line of research fundamentally guided by specific biological questions is the deployment of innovative modelling and analysis methods for the *in silico* studies.

Synthetic biology, or bioengineering, aims at designing and constructing *in vivo* biological systems that performs novel, useful tasks. This is achieved by reingeneering existing natural biological systems. While the construction of simple intracellular circuits has shown the feasibility of the approach, 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). How should cells be genetically modified such that the desired behavior robustly emerges from cell interactions? In collaboration with Dirk Drasdo (EPI BANG), we develop *abstraction methods for multiscale systems* to make the design and optimization of such systems computationally tractable and investigate the mammalian tissue homeostasis problem from a bioengineering point of view. Then, in collaboration with the Weiss lab (MIT), we construct and test *in vitro* the proposed designs in actively-growing mammalian cells.

The rational design of synthetic systems relies however on a good quantitative understanding of the functioning of the various processes involved. To acquire that knowledge, one observes the cell reaction to a range of external perturbations. However, current experimental techniques do not allow precise perturbations of cellular processes over a long time period. To make progress on this problem, we develop an experimental platform for the *closed-loop control* of intracellular processes. In collaboration with the MSC lab (CNRS/Paris Diderot U), we develop models of the controlled cellular system, generate quantitative data for parameter identification, and develop real-time control approaches. The integration of all these elements results in an original platform combining hardware (microfluidic device and microscope) and software (cell tracking and model predictive control algorithms). More specifically, by setting up an external, *in silico* feedback loop, we investigate the strengths and time scales of natural feedback loops, responsible for cell adaptation to environmental fluctuations.

CONVECS Team

3. Scientific Foundations

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 [36] 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 analyses 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 operability 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 analyses.

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 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 enable 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 catalogue of interoperable components serving as building blocks for new analysis features. A long-term goal of this approach is to provide an increasingly large catalogue 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 catalogue. 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.

DART Project-Team

3. Scientific Foundations

3.1. Introduction

The main research topic of the DaRT team-project concerns the hardware/software codesign of embedded systems with high performance processing units like DSP or SIMD processors. A special focus is put on multi processor architectures on a single chip (System-on-Chip). The contribution of DaRT is organized around the following items:

Co-modeling for High Performance SoC design: We define our own metamodels to specify application, architecture, and (software hardware) association. These metamodels present new characteristics as high level data parallel constructions, iterative dependency expression, data flow and control flow mixing, hierarchical and repetitive application and architecture models. All these metamodels are implemented with respect to the MARTE standard profile of the OMG group, which is dedicated to the modeling of embedded and real-time systems.

Model-based optimization and compilation techniques: We develop automatic transformations of data parallel constructions. They are used to map and to schedule an application on a particular architecture. This architecture is by nature heterogeneous and appropriate techniques used in the high performance community can be adapted. We developed new heuristics to minimize the power consumption. This new objective implies to specify multi criteria optimization techniques to achieve the mapping and the scheduling.

SoC simulation, verification and synthesis: We develop a SystemC based simulation environment at different abstraction levels for accurate performance estimation and for fast simulation. To address an architecture and the applications mapped on it, we simulate in SystemC at different abstraction levels the result of the SoC design. This simulation allows us to verify the adequacy of the mapping and the schedule, e.g., communication delay, load balancing, memory allocation. We also support IP (Intellectual Property) integration with different levels of specification. On the other hand, we use formal verification techniques in order to ensure the correctness of designed systems by particularly considering the synchronous approach. Finally, we transform MARTE models of data intensive algorithms in VHDL, in order to synthesize a hardware implementation.

3.2. Co-modeling for HP-SoC design

Modeling, UML, Marte, MDE, Transformation, Model, Metamodel

The main research objective is to build a set of metamodels (application, hardware architecture, association, deployment and platform specific metamodels) to support a design flow for SoC design. We use a MDE (Model Driven Engineering) based approach.

3.2.1. Foundations

3.2.1.1. System-on-Chip Design

SoC (System-on-Chip) can be considered as a particular case of embedded systems. SoC design covers a lot of different viewpoints including the application modeling by the aggregation of functional components, the assembly of existing physical components, the verification and the simulation of the modeled system, and the synthesis of a complete end-product integrated into a single chip.

The model driven engineering is appropriate to deal with the multiple abstraction levels. Indeed, a model allows several viewpoints on information defined only once and the links or transformation rules between the abstraction levels permit the re-use of the concepts for a different purpose.

3.2.1.2. Model-driven engineering

Model Driven Engineering (MDE) [68] is now recognized as a good approach for dealing with System on Chip design issues such as the quick evolution of the architectures or always growing complexity. MDE relies on the model paradigm where a model represents an abstract view of the reality. The abstraction mechanism avoids dealing with details and eases reusability.

A common MDE development process is to start from a high level of abstraction and to go to a targeted model by flowing through intermediate levels of abstraction. Usually, high level models contain only domain specific concepts, while technological concepts are introduced smoothly in the intermediate levels. The targeted levels are used for different purposes: code generation, simulation, verification, or as inputs to produce other models, etc. The clear separation between the high level models and the technological models makes it easy to switch to a new technology while re-using the previous high level designs. Transformations allow to go from one model at a given abstraction level to another model at another level, and to keep the different models synchronized

In an MDE approach, a SoC designer can use the same language to design application and architecture. Indeed, MDE is based on proved standards: UML 2 [38] for modeling, the MOF (Meta Object Facilities [63]) for metamodel expression and QVT [64] for transformation specifications. Some profiles, i.e. UML extensions, have been defined in order to express the specificities of a particular domain. In the context of embedded system, the MARTE profile in which we contribute follows the OMG standardization process.

3.2.1.3. Models of computation

We briefly present our reference models of computation that consist of the Array-OL language and the synchronous model. The former allows us to express the parallelism in applications while the latter favors the formal validation of the design.

Array-OL. The Array-OL language [52], [53], [48], [47] is a mixed graphical-textual specification language dedicated to express multidimensional intensive signal processing applications. It focuses on expressing all the potential parallelism in the applications by providing concepts to express data-parallel access in multidimensional arrays by regular tilings. It is a single assignment first-order functional language whose data structures are multidimensional arrays with potentially cyclic access.

The synchronous model. The synchronous approach [46] proposes formal concepts that favor the trusted design of embedded real-time systems. Its basic assumption is that computation and communication are instantaneous (referred to as "synchrony hypothesis"). The execution of a system is seen through the chronology and simultaneity of observed events. This is a main difference from visions where the system execution is rather considered under its chronometric aspect (i.e., duration has a significant role). There are different synchronous languages with strong mathematical foundations. These languages are associated with tool-sets that have been successfully used in several critical domains, e.g. avionics, nuclear power plants.

In the context of the DaRT project, we consider declarative languages such as Lustre [50] and Signal [61] to model various refinements of Array-OL descriptions in order to deal with the control aspect as well as the temporal aspect present in target applications. The first aspect is typically addressed by using concepts such as mode automata, which are proposed as an extension mechanism in synchronous declarative languages. The second aspect is studied by considering temporal projections of array dimensions in synchronous languages based on clock notion. The resulting synchronous models are analyzable using the formal techniques and tools provided by the synchronous technology.

3.2.2. Past contributions of the team on topics continued in 2012

The new team DaRT has been created in order to finalize the works started in the DaRT EPI, and also to explore new topics. We here remind the past contributions of the team on the topics we continued to work on during 2012.

Our proposal is partially based upon the concepts of the "Y-chart" [57]. The MDE contributes to express the model transformations which correspond to successive refinements between the abstraction levels.

Metamodeling brings a set of tools which enable us to specify our application and hardware architecture models using UML tools, to reuse functional and physical IPs, to ensure refinements between abstraction levels via mapping rules, to initiate interoperability between the different abstraction levels used in a same codesign, and to ensure the opening to other tools, like verification tools, thought the use of standards.

The application and the hardware architecture are modeled separately using similar concepts inspired by Array-OL to express the parallelism. The placement and scheduling of the application on the hardware architecture is then expressed in an association model.

All the previously defined models, application, architecture and association, are platform independent and they conform to the MARTE OMG Profil (figure 1). No component is associated with an execution, simulation or synthesis technology. Such an association targets a given technology (OpenMP, OpenCL, SystemC/PA, VHDL, etc.). Once all the components are associated with some IPs of the GasparLib library, the deployment is fully realized. This result can be transformed to further abstraction level models via some model transformations (figure 2).

The simulation results can lead to a refinement of the initial application, hardware architecture, association and deployment models. We propose a methodology to work with all these different models. The design steps are:

1. Separation of application and hardware architecture modeling.
2. Association with semi-automatic mapping and scheduling.
3. Selection of IPs from libraries for each element of application/architecture models, to achieve the deployment.
4. Automatic generation of the various platform specific simulation or execution models.
5. Automatic simulation or execution code generation with calls to the IPs.
6. Refinement at the highest level taking account of the simulation results.

3.2.2.1. High-level modeling in Gaspard2

In Gaspard2, models are described by using the recent OMG standard MARTE profile combined with a few native UML concepts and some extensions.

The new release of Gaspard2 uses different packages of MARTE for UML modeling. The Hardware Resource Model (HRM) concepts of MARTE enable to describe the hardware part of a system. The Repetitive Structure Modeling (RSM) concepts allow one to describe repetitive structures (DaRT team was the main contributor of this MARTE package definition). Finally, the Generic Component Modeling (GCM) concepts are used as the base for component modeling.

The above concepts are expressive enough to permit the modeling of different aspects of an embedded system:

- functionality (or applicative part): the focus is mainly put on the expression of data dependencies between components in order to describe an algorithm. Here, the manipulated data are mainly multidimensional arrays. Furthermore, a form of reactive control can be described in modeled applications via the notion of execution modes. This last aspect is modeled with the help of some native UML notions in addition to MARTE.
- hardware architecture: similar mechanisms are also used here to describe regular architectures in a compact way. Regular parallel computation units are more and more present in embedded systems, especially in SoCs. HRM is fully used to model these concepts. Some extensions are proposed for NoC design and FPGA specifications. The GPU have a particular memory hierarchy. In order to model the memory details, we extend the MARTE metamodel to describe low level characteristics of the memory.
- association of functionality with hardware architecture: the main issues concern the allocation of the applicative part of a system onto the available computation resources, and the scheduling. Here also, the allocation model takes advantage of the repetitive and hierarchical representation offered by MARTE to enable the association at different granularity levels, in a factorized way.

In addition to the above usual design aspects, Gaspard2 also defines a notion of *deployment* specification (see Figure 1) in order to select compilable IPs from libraries, at this time models can produce codes. The corresponding package defines concepts that (i) enable to describe the relation between a MARTE representation of an elementary component (a box with ports) to a text-based code (and Intellectual Property - IP, or a function with arguments), and (ii) allow one to inform the Gaspard2 transformations of specific behaviors of each component (such as average execution time, power consumption...) in order to generate a high abstraction level simulation in adequacy with the real system. Recently this package was extended to design reconfigurable systems using dynamical deployment.

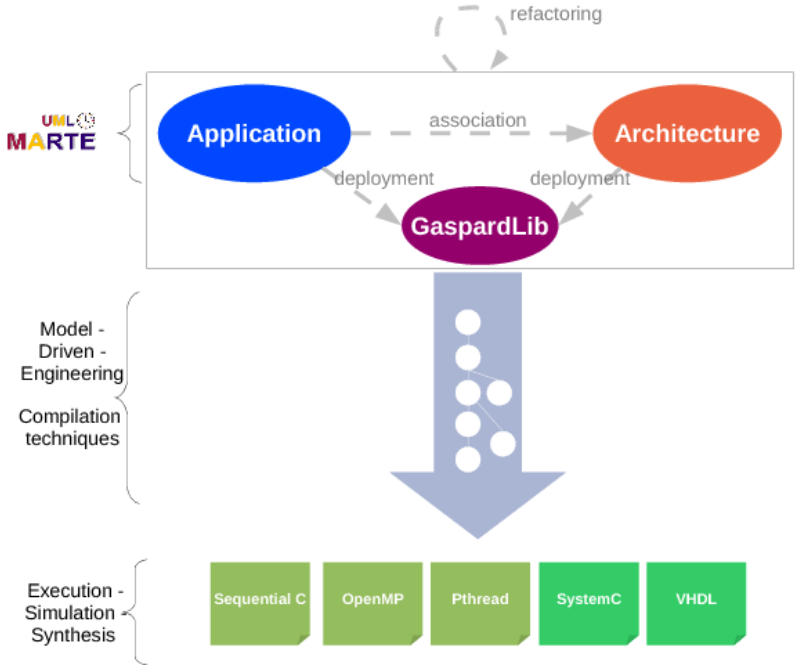


Figure 1. Overview of the design concepts.

3.2.2.2. Intermediate concept modeling and transformations

Gaspard2 targets different technologies for various purposes: formal verification, high-performance computing, simulation and hardware synthesis (Figure 1). This is achieved via model transformations that relate intermediate representations towards the final target representations.

- A metamodel for procedural language with OpenMP (OpenMP in Figure 1). It is inspired by the ANSI C and Fortran grammars and extended by OpenMP statements [41]. The aim of this metamodel is to use the same model to represent Fortran and C code. Thus, from an OpenMP model, it is possible to generate OpenMP/Fortran or OpenMP/C. The generated code includes parallelism directives and control loops to distribute task (IPs code) repetitions over processors [70].
- A VHDL metamodel (VHDL in Figure 1). It gathers the necessary concepts to describe hardware accelerators at the RTL (Register Transfer Level) level, which allows the hardware execution of applications. This metamodel introduces, *e.g.*, the notions of *clock* and *register* in order to manipulate some of the usual hardware design concepts. It is precise enough to enable the generation of synthetizable HDL code [60].

- The two metamodels SystemC and Pthread was redefined to implement both a multi-thread execution model. These are described in the " New results" part.
- Synchronous metamodel (Synchronous Equational). It was used to benefit of the verification tools of synchronous languages. It is not yet maintained in the new release of Gaspard2.

The transformation scheme. In order to target these metamodels, several transformations have been developed (Figure 2). *MartePortInstance* introduces into the MARTE metamodel the concept of *PortInstance* corresponding to an instance of port associated to a part. The *ExplicitAllocation* transformation explicits the association of each application part on the processing units, according to the association of other elements in the application hierarchy. The *LinkTopologyTask* transformation replaces the connectors between a component and an inner repeated part by a task managing the data (*TilerTask*). The scheduling of the application tasks is decomposed into three transformations, *Synchronisation* that associates, to each application component, a local graph of tasks corresponding to its parts; *GlobalSynchronization* that computes a global graph of tasks for the complete application from the local graphs of tasks; and *Scheduling* that schedules the tasks from the global graph. *TilerMapping* maps the *TilerTasks* onto processors. The management of the data in the memory is performed through two transformations. *MemoryMapping* maps the data into memory *i.e.* creates the variables and allocates address spaces. *AddressComputation* computes addresses for each variable. Finally, some transformations are dedicated to targets: *Functional* introduces the concepts relative to procedural languages. *pThread* transforms MARTE elementary tasks into threads and the connectors into buffers. *SystemC* traduces the MARTE architecture into concepts of the SystemC language.

3.2.2.3. MARTE extensions for reconfigurable based systems

Reconfigurable FPGA based Systems-on-Chip (SoC) architectures are increasingly becoming the preferred solution for implementing modern embedded systems. However due to the tremendous amount of hardware resources available in these systems, new design methodologies and tools are required to reduce their design complexity.

In previous work, we provided an initial contribution to the modeling of these systems by extending MARTE profile to incorporate significant design criteria such as power consumption.

In its current version, MARTE lacks dynamic reconfiguration concepts. Even these later are necessary to model and implement rapid prototypes for complex systems.

Our objective is to define all necessary concepts for dynamic reconfiguration issues regarding configuration latency, resources number, etc. Afterwards, these concepts will be integrated to MARTE to obtain an extended and complete profile, which can be called Reconfigurable MARTE (RecoMARTE).

Our current proposals permit us to model fine grain reconfigurable FPGA architectures with an initial extension of the MARTE profile to model Dynamic Reconfiguration at a high-level description.

Since a controller is essential for managing a dynamically reconfigurable region, we modeled a state machine at high abstraction levels using UML state machine diagrams. This state machine is responsible for switching between the available configurations.

As a future work, we will analyze the reconfigurable design flow of Xilinx from the design partitioning to the bitstream generation stage. It is a starting point for understanding how to generate configuration files. Then, we will extract relevant data to define our own design flow.

3.2.2.4. Traceability

We use the transformation mechanism to assist a tester in the mutation analysis process dedicated to model transformations. The mutation analysis aims to qualify a test model set. More precisely, errors are voluntary injected in transformation and the ability of the test models set to highlight these errors is analyzed. If the number of highlighted errors, *i.e.* if the test model set is not enough qualified, new models have to be added in order to raise the set quality [62]. Our approach relies on the hypothesis that it is easier to modify an existing model than to create a new one from scratch. The local trace, coupled to a mutation matrix, helps the tester to identify adequate test models and their relevant parts to modify in order to improve the test data set.

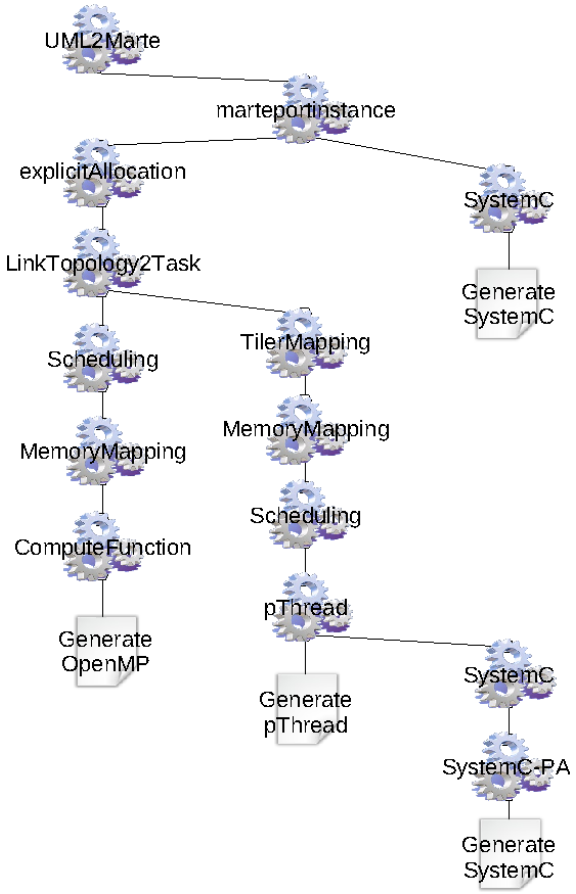


Figure 2. Overview of the transformation chains.

We propose a semi-automation approach that can automatically generate new test model in some cases and efficiently assist the testers in others cases [45].

3.3. Model-based optimization and compilation techniques

Scheduling, Mapping, Compilation, Optimization, Heuristics, Power Consumption, Data-parallelism

3.3.1. Foundations

3.3.1.1. Optimization for parallelism

We study optimization techniques to produce "good" schedules and mappings of a given application onto a hardware SoC architecture. These heuristic techniques aim at fulfilling the requirements of the application, whether they be real time, memory usage or power consumption constraints. These techniques are thus multi-objective and target heterogeneous architectures.

We aim at taking advantage of the parallelism (both data-parallelism and task parallelism) expressed in the application models in order to build efficient heuristics.

Our application model has some good properties that can be exploited by the compiler: it expresses all the potential parallelism of the application, it is an expression of data dependencies –so no dependence analysis is needed–, it is in a single assignment form and unifies the temporal and spatial dimensions of the arrays. This gives to the optimizing compiler all the information it needs and in a readily usable form.

3.3.1.2. Transformation and traceability

Model to model transformations are at the heart of the MDE approach. Anyone wishing to use MDE in its projects is sooner or later facing the question: how to perform the model transformations? The standardization process of Query View Transformation [64] was the opportunity for the development of transformation engine as Viatra, Moflon or Sitra. However, since the standard has been published, only few of investigating tools, such as ATL [1] (a transformation dedicated tool) or Kermeta [2] (a generalist tool with facilities to manipulate models) are powerful enough to execute large and complex transformations such as in the Gaspard2 framework. None of these engine is fully compliant with the QVT standard. To solve this issue, new engine relying on a subset of the standard recently emerged such as QVTO [3] and smartQVT. These engines implement the QVT Operational language.

Traceability may be used for different purposes such as understanding, capturing, tracking and verification on software artifacts during the development life cycle [58]. MDE has as main principle that everything is a model, so trace information is mainly stored as models. Solutions are proposed to keep the trace information in the initials models source or target [71]. The major drawbacks of this solution are that it pollutes the models with additional information and it requires adaptation of the metamodels in order to take into account traceability. Using a separate trace model with a specific semantics has the advantage of keeping trace information independent of initial models [59].

3.3.2. Past contributions of the team on topics continued in 2012

The new team DaRT has been created in order to finalize the works started in the DaRT EPI, and also to explore new topics. We here remind the past contributions of the team on the topics we continued to work on during 2012.

[1] http://www.eclipse.org/m2m/atl
[2] http://www.kermeta.org
[3] http://www.eclipse.org/m2m/qvto/doc

3.3.2.1. Transformation techniques

In the previous version of Gaspard2, model transformations were complex and monolithic. They were thus hardly evolvable, reusable and maintainable. We thus proposed to decompose complex transformations into smaller ones jointly working in order to build a single output model [56]. These transformations involve different parts of the same input metamodel (e.g. the MARTE metamodel); their application field is localized. The localization of the transformation was ensured by the definition of the intermediary metamodels as delta. The delta metamodel only contains the few concepts involved in the transformation (i.e. modified, or read). The specification of the transformations only uses the concepts of these deltas. We defined the Extend operator to build the complete metamodel from the delta and transposed the corresponding transformations. The complete metamodel corresponds to the merge between the delta and the MARTE metamodel or an intermediary metamodel. The transformation then becomes the chaining of metamodel shifts and the localized transformation. This way to define the model transformations has been used in the Gaspard2 environment. It allowed a better modularity and thus also reusability between the various chains.

3.3.2.2. Traceability

Our traceability solution relies on two models the Local and the Global Trace metamodels. The former is used to capture the traces between the inputs and the outputs of one transformation. The Global Trace metamodel is used to link Local Traces according to the transformation chain. The local trace also proposes an alternative "view" to the common traceability mechanism that does not refers to the execution trace of the transformation engine. It can be used whatever the used transformation language and can easily complete an existing traceability mechanism by providing a more finer grain traceability [43].

Furthermore, based on our trace metamodels, we developed algorithms to ease the model transformation debug. Based on the trace, the localization of an error is eased by reducing the search field to the sequence of the transformation rule calls [44].

3.3.2.3. Modeling for GPU

The model described in UML with Marte profile model is chained in several inout transformations that adds and/or transforms elements in the model. For adding memory allocation concepts to the model, a QVT transformation based on «Memory Allocation Metamodel» provides information to facilitate and optimize the code generation. Then a model to text transformation allows to generate the C code for GPU architecture. Before the standard releases, Acceleo is appropriate to get many aspects from the application and architecture model and transform it in CUDA (.cu, .cpp, .c, .h, Makefile) and OpenCL (.cl, .cpp, .c, .h, Makefile) files. For the code generation, it's required to take into account intrinsic characteristics of the GPUs like data distribution, contiguous memory allocation, kernels and host programs, blocks of threads, barriers and atomic functions.

3.3.2.4. GPGPU code production

The solution of large, sparse systems of linear equations « Ax=b » presents a bottleneck in sequential code executing on CPU. To solve a system bound to Maxwell's equations on Finite Element Method (FEM), a version of conjugate gradient iterative method was implemented in CUDA and OpenCL as well. The aim is to accelerate and verify the parallel code on GPUs. The first results showed a speedup around 6 times against sequential code on CPU. Another approach uses an algorithm that explores the sparse matrix storage format (by rows and by columns). This one did not increase the speedup but it allows to evaluate the impact of the access to the memory.

3.3.2.5. From MARTE to OpenCL.

We have proposed an MDE approach to generate OpenCL code. From an abstract model defined using UML/MARTE, we generate a compilable OpenCL code and then, a functional executable application. As MDE approach, the research results provide, additionally, a tool for project reuse and fast development for not necessarily experts. This approach is an effective operational code generator for the newly released OpenCL standard. Further, although experimental examples use mono device(one GPU) example, this approach provides resources to model applications running on multi devices (homogeneously configured). Moreover, we provide two main contributions for modeling with UML profile to MARTE. On the one hand, an approach to model distributed memory simple aspects, i.e. communication and memory allocations. On the other hand,

an approach for modeling the platform and execution models of OpenCL. During the development of the transformation chain, an hybrid metamodel was proposed for specifying of CPU and GPU programming models. This allows generating other target languages that conform the same memory, platform and execution models of OpenCL, such as CUDA language. Based on other created model to text templates, future works will exploit this multi language aspect. Additionally, intelligent transformations can determine optimization levels in data communication and data access. Several studies show that these optimizations increase remarkably the application performance.

3.3.2.6. Formal techniques for construction, compilation and analysis of domain-specific languages

The increasing complexity of software development requires rigorously defined *domain specific modelling languages* (DSML). Model-driven engineering (MDE) allows users to define their language's syntax in terms of *metamodels*. Several approaches for defining operational semantics of DSML have also been proposed [69], [51], [42], [49], [65]. We have also proposed one such approach, based on representing models and metamodels as algebraic specifications, and operational semantics as rewrite rules over those specifications [54], [67]. These approaches allow, in principle, for model execution and for formal analyses of the DSML. However, most of the time, the executions/analyses are performed via transformations to other languages: code generation, resp. translation to the input language of a model checker. The consequence is that the results (e.g., a program crash log, or a counterexample returned by a model checker) may not be straightforward to interpret by the users of a DSML. We have proposed in [66] a formal and operational framework for tracing such results back to the original DSML's syntax and operational semantics, and have illustrated it on SPEM, a language for timed process management.

DEDUCTEAM Team (section vide)

ESPRESSO Project-Team

3. Scientific Foundations

3.1. Introduction

Embedded systems are not new, but their pervasive introduction in ordinary-life objects (cars, telephone, home appliances) brought a new focus onto design methods for such systems. New development techniques are needed to meet the challenges of productivity in a competitive environment. Synchronous languages rely on the *synchronous hypothesis*, which lets computations and behaviors be divided into a discrete sequence of *computation steps* which are equivalently called *reactions* or *execution instants*. In itself this assumption is rather common in practical embedded system design.

But the synchronous hypothesis adds to this the fact that, *inside each instant*, the behavioral propagation is well-behaved (causal), so that the status of every signal or variable is established and defined prior to being tested or used. This criterion, which may be seen at first as an isolated technical requirement, is in fact the key point of the approach. It ensures strong semantic soundness by allowing universally recognized mathematical models to be used as supporting foundations. In turn, these models give access to a large corpus of efficient optimization, compilation, and formal verification techniques. The synchronous hypothesis also guarantees full equivalence between various levels of representation, thereby avoiding altogether the pitfalls of non-synthesizability of other similar formalisms. In that sense the synchronous hypothesis is, in our view, a major contribution to the goal of *model-based design* of embedded systems.

We shall describe the synchronous hypothesis and its mathematical background, together with a range of design techniques enpowered by the approach. Declarative formalisms implementing the synchronous hypothesis can be cast into a model of computation [8] consisting of a domain of traces or behaviors and of semi-lattice structure that renders the synchronous hypothesis using a timing equivalence relation: clock equivalence. Asynchrony [32] can be superimposed on this model by considering a flow equivalence relation as well as heterogeneous systems [33] by parameterizing composition with arbitrary timing relations.

3.2. Polychronous model of computation

We consider a partially-ordered set of tags t to denote instants seen as symbolic periods in time during which a reaction takes place. The relation $t_1 \leq t_2$ says that t_1 occurs before t_2. Its minimum is noted 0. A totally ordered set of tags C is called a *chain* and denotes the sampling of a possibly continuous or dense signal over a countable series of causally related tags. Events, signals, behaviors and processes are defined as follows:

- an *evente* is a pair consisting of a value v and a tag t,
- a *signals* is a function from a *chain* of tags to a set of values,
- a *behaviorb* is a function from a set of names x to signals,
- a *processp* is a set of behaviors that have the same domain.

In the remainder, we write $\mathrm{tags}(s)$ for the tags of a signal s, $\mathrm{vars}(b)$ for the domain of b, $b|_X$ for the projection of a behavior b on a set of names X and b/X for its complementary.

Figure 1 depicts a behavior b over three signals named x, y and z. Two frames depict timing domains formalized by chains of tags. Signals x and y belong to the same timing domain: x is a down-sampling of y. Its events are synchronous to odd occurrences of events along y and share the same tags, e.g. t_1. Even tags of y, e.g. t_2, are ordered along its chain, e.g. $t_1 < t_2$, but absent from x. Signal z belongs to a different timing domain. Its tags are not ordered with respect to the chain of y.

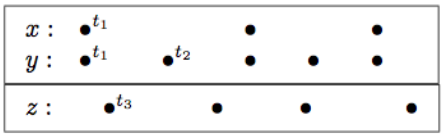


Figure 1. Behavior b over three signals x, y and z in two clock domains

3.2.1. Composition

Synchronous composition is noted $p \,|\, q$ and defined by the union $b \cup c$ of all behaviors b (from p) and c (from q) which hold the same values at the same tags $b|_I = c|_I$ for all signal $x \in I = \text{vars}(b) \cap \text{vars}(c)$ they share. Figure 2 depicts the synchronous composition (Figure 2 , right) of the behaviors b (Figure 2 , left) and the behavior c (Figure 2 , middle). The signal y, shared by b and c, carries the same tags and the same values in both b and c. Hence, $b \cup c$ defines the synchronous composition of b and c.

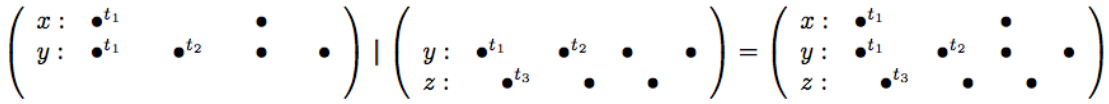


Figure 2. Synchronous composition of $b \in p$ and $c \in q$

3.2.2. Scheduling

A scheduling structure is defined to schedule the occurrence of events along signals during an instant t. A scheduling $\rightarrow$ is a pre-order relation between dates x_t where t represents the time and x the location of the event. Figure 3 depicts such a relation superimposed to the signals x and y of Figure 1 . The relation $y_{t_1} \rightarrow x_{t_1}$, for instance, requires y to be calculated before x at the instant t_1. Naturally, scheduling is contained in time: if $t < t'$ then $x_t \rightarrow^b x_{t'}$ for any x and b and if $x_t \rightarrow^b x_{t'}$ then $t' \neg < t$.

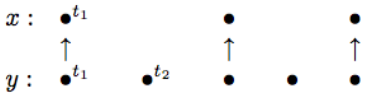


Figure 3. Scheduling relations between simultaneous events

3.2.3. Structure

A synchronous structure is defined by a semi-lattice structure to denote behaviors that have the same timing structure. The intuition behind this relation is depicted in Figure 4 . It is to consider a signal as an elastic with

ordered marks on it (tags). If the elastic is stretched, marks remain in the same relative (partial) order but have more space (time) between each other. The same holds for a set of elastics: a behavior. If elastics are equally stretched, the order between marks is unchanged.

In Figure 4 , the time scale of x and y changes but the partial timing and scheduling relations are preserved. Stretching is a partial-order relation which defines clock equivalence. Formally, a behavior c is a *stretching* of b of same domain, written $b \leq c$, iff there exists an increasing bijection on tags f that preserves the timing and scheduling relations. If so, c is the image of b by f. Last, the behaviors b and c are said *clock-equivalent*, written $b \sim c$, iff there exists a behavior d s.t. $d \leq b$ and $d \leq c$.

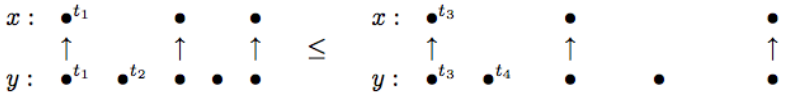


Figure 4. Relating synchronous behaviors by stretching.

3.3. A declarative design language

Signal [34], [48], [49], [41] is a declarative design language expressed within the polychronous model of computation. In Signal, a process P is an infinite loop that consists of the synchronous composition $P \,|\, Q$ of simultaneous equations $x := y\ f\ z$ over signals named x, y, z. The restriction of a signal name x to a process P is noted P/x.

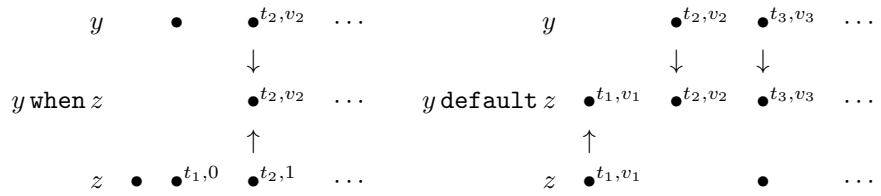
$$P, Q ::= x := y\ f\ z \mid P/x \mid P \,|\, Q$$

Equations $x := y\ f\ z$ in Signal more generally denote processes that define timing relations between input and output signals. There are four primitive combinators in Signal:

- delay $x := y \,\$\, \texttt{init}\ v$, initially defines the signal x by the value v and then by the previous value of the signal y. The signal y and its delayed copy $x = y \,\$\, \texttt{init}\ v$ are synchronous: they share the same set of tags $t_1, t_2, \cdots$. Initially, at t_1, the signal x takes the declared value v and then, at tag t_n, the value of y at tag t_{n-1}.

$$\begin{array}{rcccc} y & \bullet^{t_1,v_1} & \bullet^{t_2,v_2} & \bullet^{t_3,v_3} & \ldots \\ y \,\$\, \texttt{init}\ v & \bullet^{t_1,v} & \bullet^{t_2,v_1} & \bullet^{t_3,v_2} & \ldots \end{array}$$

- sampling $x := y\ \texttt{when}\ z$, defines x by y when z is true (and both y and z are present); x is present with the value v_2 at t_2 only if y is present with v_2 at t_2 and if z is present at t_2 with the value true. When this is the case, one needs to schedule the calculation of y and z before x, as depicted by $y_{t_2} \rightarrow x_{t_2} \longleftarrow z_{t_2}$.
- merge $x := y\ \texttt{default}\ z$, defines x by y when y is present and by z otherwise. If y is absent and z present with v_1 at t_1 then x holds (t_1, v_1). If y is present (at t_2 or t_3) then x holds its value whether z is present (at t_2) or not (at t_3).



The structuring element of a Signal specification is a process. A process accepts input signals originating from possibly different clock domains to produce output signals when needed. This allows, for instance, to specify a counter where the inputs `tick` and `reset` and the output `value` have independent clocks. The body of `counter` consists of one equation that defines the output signal `value`. Upon the event `reset`, it sets the count to 0. Otherwise, upon a `tick` event, it increments the count by referring to the previous value of `value` and adding 1 to it. Otherwise, if the count is solicited in the context of the counter process (meaning that its clock is active), the counter just returns the previous count without having to obtain a value from the `tick` and `reset` signals.

```
process counter = (? event tick, reset; ! integer value;)
    (| value := (0 when reset)
        default ((value$ init 0 + 1) when tick)
        default (value$ init 0)
    |);
```

A Signal process is a structuring element akin to a hierarchical block diagram. A process may structurally contain sub-processes. A process is a generic structuring element that can be specialized to the timing context of its call. For instance, the definition of a synchronized counter starting from the previous specification consists of its refinement with synchronization. The input tick and reset clocks expected by the process `counter` are sampled from the boolean input signals `tick` and `reset` by using the `when tick` and `when reset`†expressions. The count is then synchronized to the inputs by the equation `reset ^= tick ^= value`.

```
process synccounter = (? boolean tick, reset; ! integer value;)
    (| value := counter (when tick, when reset)
     | reset ^= tick ^= value
     |);
```

3.4. Compilation of Signal

Sequential code generation starting from a Signal specification starts with an analysis of its implicit synchronization and scheduling relations. This analysis yields the control and data-flow graphs that define the class of sequentially executable specifications and allow to generate code.

3.4.1. Synchronization and scheduling specifications

In Signal, the clock $\hat{}x$ of a signal x denotes the set of instants at which the signal x is present. It is represented by a signal that is true when x is present and that is absent otherwise. Clock expressions represent control. The clock `when` x (resp. `when not` x) represents the time tags at which a boolean signal x is present and true (resp. false).

The empty clock is written $\hat{}0$ and clock expressions e combined using conjunction, disjunction and symmetric difference. Clock equations E are Signal processes: the equation $e\hat{} = e'$ synchronizes the clocks e and e' while $e\hat{} < e'$ specifies the containment of e in e'. Explicit scheduling relations $x \rightarrow y$ `when` e allow to schedule the calculation of signals (e.g. x after y at the clock e).

$$e ::= \hat{x} \mid \texttt{when}\, x \mid \texttt{not}\, x \mid e\hat{+}e' \mid e\hat{-}e' \mid e\hat{*}e' \mid \hat{0} \quad \text{(clock expression)}$$
$$E ::= () \mid e\hat{=}e' \mid e\hat{<}e' \mid x \to y\, \texttt{when}\, e \mid E \,|\, E' \mid E/x \quad \text{(clock relations)}$$

3.4.2. Synchronization and scheduling analysis

A Signal process P corresponds to a system of clock and scheduling relations E that denotes its timing structure. It can be defined by induction on the structure of P using the inference system $P : E$ of Figure 5 .

```
x := y$ init v   : ^x ^= ^y
x := y when z    : ^x ^= ^y when z | y -> x when z
x := y default z : ^x ^= ^y default ^z | y -> x when ^y | z -> x when ^z ^- ^y
```

Figure 5. Clock inference system

3.4.3. Hierarchization

The clock and scheduling relations E of a process P define the control-flow and data-flow graphs that hold all necessary information to compile a Signal specification upon satisfaction of the property of *endochrony*. A process is said endochronous iff, given a set of input signals and flow-equivalent input behaviors, it has the capability to reconstruct a unique synchronous behavior up to clock-equivalence: the input and output signals are ordered in clock-equivalent ways.

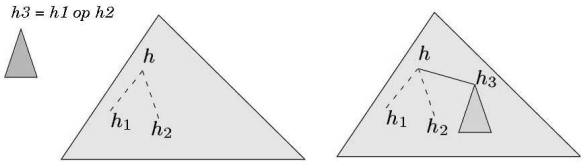


Figure 6. Hierarchization of clocks

To determine the order $x \preceq y$ in which signals are processed during the period of a reaction, clock relations E play an essential role. The process of determining this order is called hierarchization and consists of an insertion algorithm which hooks elementary control flow graphs (in the form of if-then-else structures) one to the others. Figure 6 , right, let h3 be a clock computed using h1 and h2. Let h be the head of a tree from which h1 and h2 are computed (an if-then-else), h3 is computed after h1 and h2 and placed under h [28].

FORMES Team

3. Scientific Foundations

3.1. Rewriting and Type theory

Coq [42] is one of the most popular proof assistant, in the academia and in the industry. Based on the Extended Calculus of Inductive Constructions, Coq has four kinds of basic entities: objects are used for computations (data, programs, proofs are objects); types express properties of objects; kinds categorize types by their logical structure. Coq's type checker can decide whether a given object satisfies a given type, and if a given type has a logical structure expressed by a given kind. Because it is possible to (uniformly) define inductive types such as lists, dependent types such as lists-of-length-n, parametric types such as lists-of-something, inductive properties such as $(even\ n)$ for some natural number n, etc, writing small specifications in Coq is an easy task. Writing proofs is a harder (non-automatable) task that must be done by the user with the help of tactics. We are interested in two challenges that one has to face with the development of formal proofs in Coq: the theoretical status of equality on the one hand, and the confidence one may have in Coq's proofs on the other hand. Our answer to the first challenge is CoqMTU, which isolates equality in a theory T, which must be first order, such as Presburger Arithmetic. Our answer to the second challenge is the (manual) certification of CoqMTU in Coq.

Rewriting is at the heart of proof systems such as the Extended Calculus of Constructions on which Coq is based, since mathematical proofs are made of reasonning steps, expressed by the typing rules of a given proof system, and computational steps, expressed by its rewrite rules. The certification of a proof system involves, in particular, proving three main properties of its rewrite rules: subject reduction (rewriting should preserve types), confluence (computations should be deterministic), and termination -computations must always terminate. The fact that falsity is not provable in a given proof system such as CoqMTU follows from the previous properties, while decidability of type-checking may require further work. These meta-theoretical proofs are indeed very complex, although at the same time very repetitive, depending on both the typing rules and the rewrite rules. A challenging research question here is to develop certification tools aiming at automating these proofs. Building such tools requires new results allowing to check subject-reduction, confluence and termination of higher-order calculi that are found in proof systems. Since subject-reduction is usually easy to check while consistency and decidability of type-checking follow, *in general*, from the others, confluence and termination are two very active research topics in this area. A last challenge to achieve these goals is the formalization itself of proof systems.

3.2. Verification

Model checking is an automatic formal verification technique [38]. In order to apply the technique, users have to formally specify desired properties on an abstract model of the system under verification. Model checkers will check whether the abstract model satisfies the given properties. If model checkers are able to prove or disprove the properties on the abstract model, they report the result and terminate. In practice, however, abstract models can be extremely complicated, model checkers may not conclude with reasonable computational resources.

Compositional reasoning is a way to ameliorate the complexity in abstract models [75]. Compositional reasoning tries to prove global properties on abstract models by establishing local properties on their components. If local properties on components are easier to verify, compositional reasoning can improve the capacity of model checking by local reasoning. Experiences however suggest that local reasoning may not suffice to establish global properties. It is rare that a global property can be established without considering their interactions. In assume-guarantee reasoning, model checkers try to verify local properties under a contextual assumption of each component. If contextual assumptions faithfully capture interactions among components, model checkers can conclude the verification of global properties.

Finding contextual assumptions however is difficult and may require clairvoyance. Interestingly, a fully automated technique for computing contextual assumptions was proposed in [41]. The automated technique formalizes the contextual assumption generation problem as a learning problem. If properties and abstract models are formalized as finite automata, then a contextual assumption is nothing but an unknown finite automaton that characterizes the environment. Applying a learning algorithm for finite automata, the automated technique will generate contextual assumptions for assume-guarantee reasoning. Experimental results show that the automated technique can outperform a monolithic and explicit verification algorithm.

The success of the learning-based assume-guarantee reasoning is however not satisfactory. Most verification tools are using implicit algorithms. In fact, implicit representations such as Binary Decision Diagrams can improve the capacity of model checking algorithms in several order of magnitudes. Early learning-based techniques, on the other hand, are based on the L^* learning algorithm using explicit representations. If a contextual assumption requires hundreds of states, the learning algorithm will take too much time to infer an assumption. Subsequently, early learning-based techniques cannot compete with monolithic implicit verification [40].

Recently, we propose assume-guarantee reasoning with implicit learning [37]. Our idea is to adopt an implicit representation used in the learning-based framework. Instead of enumerating states of contextual assumptions explicitly, our new technique computes transition relations as an implicit representation of contextual assumptions. Using a learning algorithm for Boolean functions, the new technique can easily compute contextual assumptions with thousands of states. Our preliminary experimental results show that the implicit learning technique can outperform interpolation-based monolithic implicit model checking in several parametrized test cases such as synchronous bus arbiters and the MSI cache coherence protocol.

Learning Boolean functions can also be applied to loop invariant inference [53], [54]. Suppose that a programmer annotates a loop with pre- and post-conditions. We would like to compute a loop invariant to verify that the annotated loop conforms to its specification. Finding loop invariants manually is very tedious. One makes a first guess and then iteratively refines the guess by examining the loop body. This process is in fact very similar to learning an unknown formula. Applying predicate abstraction and decision procedures, a learning algorithm for Boolean functions can infer loop invariants generated by a given set of atomic predicates. Preliminary experimental results show that the learning-based technique is effective for annotated loops extracted from source codes of Linux and SPEC2000 benchmarks.

Although implicit learning techniques have been developed for assume-guarantee reasoning and loop invariant inference successfully, challenges still remain. Currently, the learning algorithm is able to infer Boolean functions over tens of Boolean variables. Contextual assumptions over tens of Boolean variables are not enough. Ideally, one would like to have contextual assumptions over hundreds (even thousands) of Boolean variables. On the other hand, it is known that learning arbitrary Boolean functions is infeasible. The scalability of implicit learning techniques cannot be improved satisfactorily by tuning the learning algorithm alone. Combining implicit learning with abstraction will be essential to improve its scalability.

Our second challenge is to extend learning-based techniques to other computation models. In addition to finite automata, probabilistic automata and timed automata are also widely used to specify abstract models. Their verification problems are much more difficult than those for finite automata. Compositional reasoning thus can improve the capacity of model checkers more significantly. Recently, the L^* algorithm is applied in assume-guarantee reasoning for probabilistic automata [46]. The new technique is unfortunately incomplete. Developing a complete learning-based assume-guarantee reasoning technique for probabilistic automata and timed automata will be very useful to their verification.

Through predicate abstraction, learning Boolean functions can be very useful in program analysis. We have successfully applied algorithmic learning to infer both quantified and quantifier-free loop invariants for annotated loops. Applying algorithmic learning to static analysis or program testing will be our last challenge. In the context of program analysis, scalability of the learning algorithm is less of an issue. Formulas over tens of atomic predicates usually suffice to characterize relation among program variables. On the other hand, learning algorithms require oracles to answer queries or generate samples. Designing such oracles necessarily

requires information extracted from program texts. How to extract information will be essential to applying algorithmic learning in static analysis or program testing.

3.3. Decision Procedures

Decision procedures are of utmost importance for us, since they are at the heart of theorem proving and verification. Research in decision procedures started several decades ago, and are now commonly used both in the academia and industry. A decision procedure [55] is an algorithm which returns a correct yes/no answer to a given input decision problem. Many real-world problems can be reduced to the decision problems, making this technique very practical. For example, Intel and AMD are developing solvers for their circuit verification tools, while Microsoft is developing decision procedures for their code analysis tools.

Mathematical logic is the appropriate tool to formulate a decision problem. Most decision problems are formulated as a decidable fragment of a first-order logic interpreted in some specific domain. On such, easy and popular fragment, is propositional (or Boolean) logic, which corresponding decision procedure is called SAT. Representing real problems in SAT often results in awkward encodings that destroy the logical structure of the original problem.

A very popular, effective recent trend is Satisfiability Modulo Theories (SMT) [74], a general technique to solve decision problems formulated as propositional formulas operating on atoms in a given background theory, for example linear real arithmetic. Existing approaches for solving SMT problems can be classified into two categories: *lazy* method [67], and *eager* method [68]. The eager method encodes an SMT problem into an equi-satisfiable SAT problem, while the lazy method employs different theory solvers for each theory and coordinates them appropriately. The eager method does allow the user to express her problem in a natural way, but does not exploit its logical structure to speed up the computation. The lazy approach is more appealing, and has prompted much interest in algorithms for the various background theories important in practice.

Our SMT solver aCiNO is based on the lazy approach. So far, it provides with two (popular) theories only: linear real arithmetic (LRA) and uninterpreted functions (UF). For efficiency consideration, the solver is implemented in an incremental way. It also invokes an online SAT solver, which is now a modified DPLL procedure, so that recovery from conflicts is possible. Our challenge here is twofold: first, to add other theories of interest for the project, we are currently working on fragments of the theory of arrays [61], [34]. The theory of arrays is important because of its use for expressing loop invariants in programs with arrays, but its full first-order theory is undecidable. We are also interested in the theory of bit vectors, very much used for hardware verification.

Theory solvers implement state-of-the-art algorithms which sophistication makes their correct implementation a delicate task. Moreover, SMT solvers themselves employ a quite complex machinery, making them error prone as well [4] We therefore strongly believe that decision procedures, and SMT provers, should come along with a formal assessment of their correctness. As usual, there are two ways: ensure the correctness of an arbitrary output by proving the code, or deliver for each input a certificate ensuring the correctness of the corresponding output when the checker says so. Developing concise certificates together with efficient certificate checkers for the various decision procedures of interest and their combination with SMT is yet another challenge which is at the heart of the project FORMES.

3.4. Simulation

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.

[4] It took almost 20 years to have a correct implementation of a correct version of Shostak's algorithm for combining decision procedures, which can be seen as an ancestor of SMT.

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, the 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 which can be implemented by software or by using a FPGA [60] simulator. These solutions tend to present slow iteration design cycles and implementing the FPGA means the hardware has already been designed at low level, which comes normally late in the project and become very costly when using large FPGA platforms. Others have implemented a co-simulation environment, using two separate technologies, typically one using a HDL and another one using an ISS [47], [50], [66]. 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 (not using FPGA) simulation environment named SimSoC, which uses a single simulation loop thanks to Transaction Level Modeling (TLM) [36], [23] 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 [35]. Another technique to implement a fast ISS is *dynamic translation* [39], [65], [44] which has been favored by many [63], [44], [64], [65] in the past decade.

With dynamic translation, the binary target instructions are fetched from memory at run-time, like in interpretive simulation. They are decoded on the first execution and the simulator translates these instructions into another representation which is stored into a cache. On further execution of the same instructions, the translated cached version is used. Dynamic translation introduces a translation time phase as part of the overall simulation time. But as the resulting cached code is re-used, the translation time is amortized over time. If the code is modified during run-time, the simulator must invalidate the cached representation. Dynamic translation provides much faster simulation while keeping the advantage of interpretive simulation as it supports the simulation of programs that have either dynamic loading or self-modifying code.

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. Both have pros and cons.

Processor simulation is also achieved in Virtual Machines such as QEMU [28] and GXEMUL [49] that emulate to a large extent the behavior of a particular hardware platform. The technique used in QEMU is a form of dynamic translation. The target code is translated directly into machine code using some pre-determined code patterns that have been pre-compiled with the C compiler. Both QEMU and GXEMUL include many device models of open-source C code, but this code is hard to reuse. The functions that emulate device accesses do not have the same profile. The scheduling process of the parallel hardware entities is not specified well enough to

guarantee the compatibility between several emulators or re-usability of third-party models using the standards from the electronics industry (e.g. IEEE 1666).

A challenge in the development of high performance simulators is to maintain simultaneously fast speed and simulation accuracy. In the FORMES 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 FORMES simulation is to extract information from the simulated applications to prove properties. Running a simulation is exercising a test case. In most cases, if a test is failing, a bug has been found. One can use model checking tools to generate tests that can be run on the simulator to check whether the test fails or not on the real application. It is also a goal of FORMES simulation activity to use such formal methods tools to detect bugs, either by generating tests, or by using formal methods tools to analyze the results of simulation sessions.

3.5. Trustworthy Software

Since the early days of software development, computer scientists have been interested in designing methods for improving software quality. Formal methods based on model checking, correctness proofs, common criteria certification, all address this issue in their own way. None of these methods, however, considers the trustworthiness of a given software system as a system-level property, requiring to grasp a given software within its environment of execution.

The major challenge we want to address here is to provide a framework in which to formalize the notion of trustworthiness, to evaluate the trustworthiness of a given software, and if necessary improve it.

To make trustworthiness a fruitful concept, our vision is to formalize it via a hierarchy of observability and controllability degrees: the more the software is observable and controllable, the more its behaviors can be trusted by users. On the other hand, users from different application domains have different expectations from the software they use. For example, aerospace embedded software should be safety-critical while e-commerce software should be insensitive to attacks. As a result, trustworthiness should be domain-specific.

A main challenge is the evaluation of trustworthiness. We believe that users should be responsible for describing the level of trustworthiness they need, in the form of formal requirements that the software should satisfy. A major issue is to come up with some predefined levels of trustworthiness for the major applicative areas. Another is to use stepwise refinement techniques to achieve the appropriate level of trustworthiness. These levels would then drive the design and implementation of a software system: the objective would be to design a model with enough details (observability) to make it possible to check all requirements of that level.

The other challenge is the effective integration of results obtained from different verification methods. There are many verification techniques, like simulation, testing, model checking and theorem proving. These methods may operate on different models of the software to be then executed, while trustworthiness should measure our trust in the real software running in its real execution environment. There are also monitoring and analysis techniques to capture the characteristics of actual executions of the system. Integrating all the analysis in order to decide the trustworthiness level of a software is quite a hard task.

GALAAD Project-Team

3. Scientific Foundations

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 complexes 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 context, 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. Scientific Foundations

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 [36]. Through collaborations, we also gained experience with several domain-specific languages based on a functional core, including XML processing (XDuce, CDuce), reactive functional programming, distributed programming (JoCaml), and hardware modeling (ReFLect).

3.2. Type systems

Type systems [49] 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 [46], [42], 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 [53], 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 [50].

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. Finally, we also participate in the FoCaLiZe project, which designs and implements an environment for combined programming and proving [23] [52].

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.

GEOMETRICA Project-Team

3. Scientific Foundations

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 makes 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 Team (section vide)

LFANT Project-Team

3. Scientific Foundations

3.1. Number fields, class groups and other invariants

Participants: Bill Allombert, Athanasios Angelakis, Karim Belabas, Julio Brau, Jean-Paul Cerri, Henri Cohen, Jean-Marc Couveignes, Andreas Enge, 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 [6]. 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 [36] 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, Jean-Marc Couveignes, Andreas Enge, Nicolas Mascot, Jérôme Milan, Damien Robert, Vincent Verneuil.

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, 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 [41], for more background material, [40]. 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. Scientific Foundations

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. Second, 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). Thus, 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. Scientific Foundations

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, César Rodríguez, 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 [98], [99], [103], 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 [83].

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 [99], [84], 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 [96], [110]). 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 [82], [87]). 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 [116], [117]. Automatizing such procedures for the supervision and management of distributed and locally monitored asynchronous systems is a mid-term goal of *MExICo*.

3.1.3. Verification of Concurrent Recursive Programs

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

3.1.3.1. Contextual nets

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 [97].

However, Petri nets are inadequate to model concurrent read accesses to the same resource. Such situations arise naturally in many circumstances, 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, which promises to yield more efficient analysis procedures. In order to realize such procedures, we shall study contextual nets and their properties, in particular the efficient construction and analysis of their unfoldings, and their applications in verification, diagnosis, and planning.

3.1.3.2. Concurrent Recursive Programs

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.

3.1.4. Testing

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

3.1.4.1. 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.

3.1.4.2. 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 [105], [88] 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.

3.1.4.3. Local Testing

Message-Sequence-Charts (MSCs) provide models of behaviors of distributed systems with communicating processes. An important problem is to test whether an implementation conforms to a specification given for instance by an HMSC. In *local testing*, one proceeds by injecting messages to the local processes and observing the responses: for each process p, a local observer records the sequence of events at p. If each local observation is consistent with some MSC defined by the specification, the implementation passes the test. If local testing on individual processes suffices to check conformance, the given specification (an HMSC language) is called locally testable. Local testability turns out to be undecidable even for regular HMSC languages [82]; the main difficulty lies in the existence of implied scenarios, i.e., global behaviors which are locally consistent with different specification scenarios. There are two approaches to attack the problem of local testing in light of this bottleneck. One is to allow joint observations of tuples of processes. This gives rise to the problem of k-testability where one allows joint observations of up to k processes [87]. We will look for structural conditions on the model or the specification ensuring k-testability. Another tactic would be to recognize that practical implementations always work with bounded buffers and impose an upper bound B on the buffer size. The set of B-bounded MSCs in the k-closure of a regular MSC language is again regular, so the B-bounded k-testability problem is decidable for all regular HMSC-definable specifications. The focus could now be on efficiently identifying the smallest k for which an HMSC specification is k-testable. Another interesting problem is to identify a minimal set of tests to validate a k-testable specification.

3.1.4.4. Goals

The first step that should be reached in the near future is the completion of asynchronous testing in the setting without any big-step synchronization. In parallel, work on the problems in local testing should progress

sufficiently to allow, in a mid-term perspective, to understand the relations and possible interconnections between local (i.e. distributed) and asynchronous (centralized) testing. This is the objective of the *TECSTES* project (2011-2014), funded by a DIGITEO *DIM/LSC* grant, and which involves 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.

In [69] and a subsequent journal submission in preparation, we develop complete testing framework for concurrent systems was developped, 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. Finally, an algorithm to construct a test suite with such properties was proposed. These result are summarized in a paper that is being prepared for a journal submission.

The mid-to long term goal (perhaps not yet to achieve in this four-year term) 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 [95] 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 [114], [108], [113], [100], [102].

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 [101],[25]. 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: Benedikt Bollig, 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.

3.3. Management of Quantitative Behavior

Participants: Sandie Balaguer, 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.

3.3.2.1. Non-sequential probabilistic processes

Practical fault diagnosis requires to select explanations of *maximal likelihood*; this leads therefore to the question what the probability of a given partially ordered execution is. In Benveniste et al. [86], [79], 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 [104], 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.

Recently, we started a new line of research with Anne Bouillard, Sidney Rosario, and Albert Benveniste in the DistribCom team at Inria Rennes, studying the likelihood of occurrence of non-sequential runs under random durations in a stochastic Petri net setting.

Once the properties of the probability measures thus obtained are understood, it will be interesting to relate them with the two above models in logical time, and understand their differences. Another mid-term goal, in parallel, is the transfer to *diagnosis*.

3.3.2.2. Distributed Markov Decision Processes

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 cooperate on this subject with Eric Fabre in the DistribCom team at Inria Rennes, in the context of the DISC project.

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 [85].
- 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 [94].

- 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 [120] .

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), (2) we want to adapt methods for discrete-event systems that require some theoretical developments in the stochastic framework and, (3) we plane 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 [111] [90]. 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. Distributed timed systems with independently evolving clocks

Participants: Benedikt Bollig, Paul Gastin.

A first step towards formal models of distributed timed systems with independently evolving clocks was done in [80]. As the precise evolution of local clock rates is often too complex or even unknown, the authors study different semantics of a given system: The *existential semantics* exhibits all those behaviors that are possible under *some* time evolution. The *universal semantics* captures only those behaviors that are possible under *all* time evolutions. While emptiness and universality of the universal semantics are in general undecidable, the existential semantics is always regular and offers a way to check a given system against safety properties. A decidable under-approximation of the universal semantics, called *reactive semantics*, is introduced to check a system for liveness properties. It assumes the existence of a *global* controller that allows the system to react upon local time evolutions. A short term goal is to investigate a *distributed* reactive semantics where controllers are located at processes and only have local views of the system behaviors.

Several questions, however, have not yet been tackled in this previous work or remain open. In particular, we plan to exploit the power of synchronization via local clocks and to investigate the *synthesis problem*: For which (global) specifications $\mathcal{S}$ can we generate a distributed timed system with independently evolving clocks $\mathcal{A}$ (over some given system architecture) such that both the reactive and the existential semantics of $\mathcal{A}$ are precisely (the semantics of) $\mathcal{S}$? In this context, it will be favorable to have partial-order based specification languages and a partial-order semantics for distributed timed systems. The fact that clocks are not shared may allow us to apply partial-order–reduction techniques.

If, on the other hand, a system is already given and complemented with a specification, then one is usually interested in controlling the system in such a way that it meets its specification. The interaction between the actual *system* and the *environment* (i.e., the local time evolution) can now be understood as a 2-player game: the system's goal is to guarantee a behavior that conforms with the specification, while the environment aims at violating the specification. Thus, building a controller of a system actually amounts to computing winning strategies in imperfect-information games with infinitely many states where the unknown or unpredictable evolution of time reflects an imperfect information of the environment. Only few efforts have been made to tackle those kinds of games. One reason might be that, in the presence of imperfect information and infinitely many states, one is quickly confronted with undecidability of basic decision problems.

3.3.4.2. Implementation of Real-Time Concurrent Systems

Participants: Sandie Balaguer, Thomas Chatain, Stefan Haar, Serge Haddad.

This is one of the tasks of the ANR ImpRo.

The objective is to provide formal guarantees on the implementation of real-time distributed systems, despite the semantic differences between the model and the code. We consider two kinds of timed models: time Petri nets [112] and networks of timed automata [81].

Time Petri Nets allow the designer to explicit concurrent parts of the system, but without having decided yet to localize the different actions on the different components. In that sense, TPNs are more abstract than networks of timed automata, which can be seen as possible (ideal) distributed implementations. This raises the question of semantical comparison of these two models in the light of preserving the maximum of concurrency.

In order to implement our models on distributed architectures, we need a way to evaluate how much the implementation preserves the concurrency that is described in the model. For this we must be able to identify concurrency in the behavior of the models. This is done by equipping the models with a concurrent semantics (unfoldings), allowing us to consider the behaviors as partial orders.

For instance, we would like to be able to transform a time Petri net into a network of timed automata, which is closer to the implementation since the processes are well identified. But we require that this transformation preserves concurrency. Yet the first works about formal comparisons of the expressivity of these models [92], [89], [91], [93], [118] do not consider preservation of concurrency.

In contrast, 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, Benjamin Monmege.

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 [78]. 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 [12]. That paper, however does not give final solutions to all the aforementioned problems. It identifies directions for future research that we will be tackling.

MUTANT Project-Team

3. Scientific Foundations

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 [20] with impacts both on signal processing using machine learning techniques and their application in real-world scenarios.

The team-project will focus on two aspects of realtime machine listening:

1. **Application-Driven Approach:** First, to enhance and foster existing application-driven approaches within the team such as realtime alignment algorithms and polyphonic pitch transcription. Our contributions on this line correspond to extensions of existing algorithmic approaches to realtime audio alignment and transcription to create new interactive application paradigms with new algorithmic

approaches. Arshia Cont's ongoing realtime alignment in *Antescofo* as well as realtime transcription using non-negative factorization methods [20] are examples of this.

2. **Music Information Geometry:** In parallel to concrete applications, we hope to theoretically contribute to the problem of signal representations of audio streams for effortless retrieval of high-level information structures. We have recently shown in [4] that the gap between the symbolic/semantic and signal aspects of music information mostly lies on constructing a well-behaved representational space before any algorithmic considerations, by employing the emerging methods of *information geometry*. Arnaud Dessein's ongoing PhD thesis is focused on this aspect of the project.

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 [28] and subsequents [31], 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 three consecutive lines:

- **Development of a Synchronous DSL for Real Time Musician-Computer Interaction:** Ongoing and continuous extensions of the *Antescofo* language following user requests and by inscribing them within a coherent framework for the handling of temporal musical relationships. José Echeveste's ongoing PhD thesis focuses on the research and development of these aspects. Recent formalizations of the *Antescofo* language has been published in [6].
- **Formal Methods:** Failure during an artistic performance should be avoided. This naturally leads to the use of formal methods, like static analysis or model checking, to ensure formally that the execution of an *Antescofo* program will satisfy some expected property. The checked properties may also provide some assistance to the composer especially in the context of "non deterministic score" in an interactive framework.

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 [35]. 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 Dr. 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" [30].

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. Scientific Foundations

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 (see the `/etc/sendmail.cf` file in the configuration of the mail system) and the transition rules describing the behaviors of tree automata. Rewriting is used in semantics in order to describe the meaning of programming languages [28] as well as in program transformations like, for example, re-engineering of Cobol programs [36]. 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 [24], a theorem prover [26] or a constraint solver [27]. 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 [29], [20], [18]. We have introduced and studied the notion of *strategy* [19], 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 [34], [31], [25] 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 [22]. 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. Scientific Foundations

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 [19], 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, [1][2] an ML extension of the synchronous language LUSTRE with higher-order features, dedicated type systems (clock calculus as a type system [19], [29], initialization analysis [30] and causality analysis [31]). 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 [28], [27], and modular code generation [20], [17].

In 2000, Marc Pouzet started to collaborate with the SCADE team of Esterel-Technologies on the design of a new version of SCADE. [3] 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. [4]

In parallel, the goal of REACTIVEML[5] 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 [42], [38] and is ongoing. The language extends OCAML[6] with Esterel-like synchronous primitives — synchronous composition, broadcast communication, pre-emption/suspension — applying the solution of Boussinot [18] to solve causality issues.

[1] http://www.di.ens.fr/~pouzet/lucid-synchrone
[2] The name is a reference to Lustre which stands for "Lucid Synchrone et Temps réel".
[3] http://www.esterel-technologies.com/products/scade-suite/
[4] http://www.3ds.com/products/catia/portfolio/dymola/overview/
[5] http://rml.lri.fr/
[6] 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 [43].

REACTIVEML has been used for simulating routing protocols in ad-hoc networks [37] and large scale sensor networks [53]. 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 [52]. 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 [21], [22]. 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 [57]. It was algorithmically expensive and limited to periodic systems. In [25], 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 [39], as was a formalization of the theory in COQ [26]. 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 [44].

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 [47] (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).

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 [32]. 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 [36], [35], [54], [33] 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 [23], [34], [50], [49]. 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 [24], [23], we ported these techniques to the GCC compiler [48], [56], [55], 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 [50], [49], [51]. 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).

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 [45], [46]. 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. Scientific Foundations

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 [44], and normalization in different logics can justify the design of new and different functional programming languages [30].
- *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 [49], and different proof search strategies can be used to justify the design of new and different logic programming languages [48].

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 [32] 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 [56] 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 [53] 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 [54], 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. Design of two level-logic systems

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 has been published is a series of papers by McDowell and Miller [46] [45], Tiu and Miller [51] [57], and Gacek, Miller, and Nadathur [2] [38]. Besides presenting formal properties of these logic, these papers also documented a number of examples where this logic demonstrated superior approaches to reasoning about a number of complex formal systems, ranging from programming languages to the λ-calculus and π-calculus.

The team has also been working on three different prototype theorem proving system that are all related to this stronger logic. These 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.
- Tac, which is a sophisticate tactic for automatically completing simple proofs involving induction and unfolding.

We are now in the process of attempting to make all of these system communicate properly. Given that these systems have been authored by different team members at different times and for different reasons, they do not formally share the same notions of syntax and proof. We are now working to revisit all of these systems and revise them so that they all work on the *same* logic and so that they can share their proofs with each other.

Currently, Chaudhuri, Miller, and Accattoli are working with our technical staff member, Heath, to redesign and restructure these systems so that they can cooperate in building proofs.

3.3. Making the case for proof certificates

The team has been considering how it might be possible to define a universal format for proofs so that any existing theorem provers can have its proofs trusted by any other prover. This is a rather ambitious project and involves a great deal of work at the infrastructure level of computational logic. As a result, we have put significant energies into considering the high-level objectives and consequences of deploying such proof certificates.

Our current thinking on this point is roughly the following. Proofs, both formal and informal, are documents that are intended to circulate within societies of humans and machines distributed across time and space in order to provide trust. Such trust might lead a mathematician to accept a certain statement as true or it might help convince a consumer that a certain software system is secure. Using this general definition of proof, we have re-examined a range of perspectives about proofs and their roles within mathematics and computer science that often appears contradictory.

Given this view of proofs as both document and object, that need to be communicated and checked, we have attempted to define a particular approach to a *broad spectrum proof certificate* format that is intended as a universal language for communicating formal proofs among computational logic systems. We identify four desiderata for such proof certificates: they must be

1. checkable by simple proof checkers,
2. flexible enough that existing provers can conveniently produce such certificates from their internal evidence of proof,
3. directly related to proof formalisms used within the structural proof theory literature, and
4. permit certificates to elide some proof information with the expectation that a proof checker can reconstruct the missing information using bounded and structured proof search.

We consider various consequences of these desiderata, including how they can mix computation and deduction and what they mean for the establishment of marketplaces and libraries of proofs. More specifics can be found in Miller's papers [8] and [47].

3.4. Combining Classical and Intuitionistic Proof Systems

In order to develop an approach to proof certificates that is as comprehensive as possible, one needs to handle theorems and proofs in both classical logic and intuitionistic logic. Yet, building two separate libraries, one for each logic, can be inconvenient and error-prone. An ideal approach would be to design a single proof system in which both classical and intuitionistic proofs can exist together. Such a proof system should allow cut-elimination to take place and should have a sensible semantic framework.

Liang and Miller have recently been working on exactly that problem. In their paper [7], they showed how to describe a general setting for specifying proofs in intuitionistic and classical logic and to achieve one framework for describing initial-elimination and cut-elimination for such these two logics. That framework allowed for some mixing of classical and intuitionistic features in one logic. A more ambitious merging of these logics was provided in their work on "polarized intuitionistic logic" in which classical and intuitionistic connectives can be used within the same formulas [14].

3.5. Deep inference

Deep inference [40], [42] 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.

3.6. 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 [39] for linear logic but also in Robinson's proof nets [55] 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., [4], [3] and [41]. 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 [4] 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. Scientific Foundations

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 [49] 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 [43], then by Howard and de Bruijn at the end of the 60's [56], [73], 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 [40] stemmed out – a formalism that is both a logic and a programming language and that is at the source of the Coq system [39].

3.1.2. Towards the calculus of constructions

The λ-calculus, defined by Church [38], 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 [34].

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 [62].

In 1985, Coquand and Huet [40], [41] in the Formel team of Inria-Rocquencourt explored an alternative approach based on Girard-Reynolds' system F [50], [68]. 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 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 [42] 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 globally moved to Futurs with a bilocalisation on Orsay and Palaiseau. 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 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 (generally first-order ones) exist.

3.3.2. Libraries

Developing libraries for programming languages takes time and generally benefits of a critical mass effect. An advantage is given to languages that start from well-established existing frameworks for which a large panel of libraries exist. Coq is such a framework.

3.4. Around and beyond the Curry-Howard correspondence

For two decades, the Curry-Howard correspondence was limited to the intuitionistic case but in 1990, an important stimulus spurred on the community following the discovery by Griffin that the correspondence was extensible to classical logic. The community then started to investigate unexplored potential fields of connection between computer science and logic. One of these fields was the computational understanding of Gentzen's sequent calculus while another one was 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 [51] that some operators known as control operators were typable by the principle of double negation elimination ($\neg\neg A \Rightarrow A$), a principle which provides classical logic.

Control operators are operators used to jump from one place of a program to another place. They were first considered in the 60's by Landin [61] and Reynolds [67] and started to be studied in an abstract way in the 80's by Felleisen *et al* [45], culminating in Parigot's $\lambda\mu$-calculus [64], a reference calculus that is in fine Curry-Howard correspondence with classical natural deduction. In this respect, control operators are fundamental pieces of the 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 and two kinds of interpretations of these rules 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 rule. This line of work culminated in 2000 with 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 [60] and Krivine's abstract machine for call-by-name evaluation [59], [57]. 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 [46]: 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. Scientific Foundations

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 is 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, Pierre-Jean Spaenlehauer, Chenqi Mou, Jules Svartz, Louise Huot, Thibault Verron.

Efficient algorithms F_4/F_5 [1] 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.

[1] 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, Pierre-Jean Spaenlehauer, Aurélien Greuet, Simone Naldi.

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, F. Martani.

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 theses 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 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 bases is a key step toward the resolution of difficult problems in this domain [2]. Using existing resolution methods is simply not enough to solve relevant problems. The main algorithmic lock 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 [3] 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 de Portzamparc, Rina Zeitoun.

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*.

[2] 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
[3] 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.

POP ART Project-Team

3. Scientific Foundations

3.1. Embedded systems and their safe design

3.1.1. Safe Design of Embedded Real-time Control Systems

The context of our work is the area of embedded real-time control systems, at the intersection between control theory and computer science. Our contribution consists of methods and tools for their safe design. The systems we consider are intrinsically safety-critical because of the interaction between the embedded, computerized controller, and a physical process having its own dynamics. Such systems are known under various names, notably *cyberphysical systems* and *embedded control systems*. What is important is to design and to analyze the safe behavior of the whole system, which introduces an inherent complexity. This is even more crucial in the case of systems whose malfunction can have catastrophic consequences, for example in transport systems (avionics, railways, automotive), production, medical, or energy production systems (nuclear).

Therefore, there is a need for methods and tools for the design of safe systems. The definition of adequate mathematical models of the behavior of the systems allows the definition of formal calculi. They in turn form a basis for the construction of algorithms for the analysis, but also for the transformation of specifications towards an implementation. They can then be implemented in software environments made available to the users. A necessary complement is the setting-up of software engineering, programming, modeling, and validation methodologies. The motivation of these problems is at the origin of significant research activity, internationally and, in particular, in the European IST network of excellence ARTISTDESIGN (Advanced Real-Time Systems).

3.1.2. Models, Methods and Techniques

The state of the art upon which we base our contributions is twofold.

From the point of view of discrete control, there is a set of theoretical results and tools, in particular in the synchronous approach, often founded on finite or infinite labeled transition systems [31], [39]. During the past years, methodologies for the formal verification [70], [41], control synthesis [72] and compilation, as well as extensions to timed and hybrid systems [69], [32] have been developed. Asynchronous models consider the interleaving of events or messages, and are often applied in the field of telecommunications, in particular for the study of protocols.

From the point of view of verification, we use the methods and tools of symbolic model-checking and of abstract interpretation. From symbolic model-checking, we use BDD techniques [37] for manipulating Boolean functions and sets, and their MTBDD extension for more general functions. Abstract interpretation [43] is used to formalize complex static analysis, in particular when one wants to analyze the possible values of variables and pointers of a program. Abstract interpretation is a theory of approximate solving of fix-point equations applied to program analysis. Most program analysis problems, among which reachability analysis, come down to solving a fix-point equation on the state space of the program. The exact computation of such an equation is generally not possible for undecidability (or complexity) reasons. The fundamental principles of abstract interpretation are: (i) to substitute to the state-space of the program a simpler domain and to transpose the equation accordingly (static approximation); and (ii) to use extrapolation (widening) to force the convergence of the iterative computation of the fix-point in a finite number of steps (dynamic approximation). Examples of static analyses based on abstract interpretation are linear relation analysis [44] and shape analysis [40].

The synchronous approach [6] [60], [61] to reactive systems design gave birth to complete programming environments, with languages like ARGOS, LUSTRE[7], ESTEREL[8], SIGNAL/ POLYCHRONY[9], LUCIDSYNCHRONE[10], SYNDEX[11], or Mode Automata. This approach is characterized by the fact that it considers periodically sampled systems whose global steps can, by synchronous composition, encompass a set of events (known as simultaneous) on the resulting transition. Generally speaking, formal methods are often used for analysis and verification; they are much less often integrated into the compilation or generation of executives (in the sense of executables of tasks combined with the host real-time operating system). They are notoriously difficult to use by end-users, who are usually experts in the application domain, not in formal techniques. This is why encapsulating formal techniques into an automated framework can dramatically improve their diffusion, acceptance, and hence impact. Our work is precisely oriented towards this direction.

3.2. Issues in Design Automation for Complex Systems

3.2.1. Hard Problems

The design of safe real-time control systems is difficult due to various issues, among them their complexity in terms of the number of interacting components, their parallelism, the difference of the considered time scales (continuous or discrete), and the distance between the various theoretical concepts and results that allow the study of different aspects of their behaviors, and the design of controllers.

A currently very active research direction focuses on the models and techniques that allow the automatic use of formal methods. In the field of verification, this concerns in particular the technique of model checking. The verification takes place after the design phase, and requires, in case of problematic diagnostics, expensive backtracks on the specification. We want to provide a more constructive use of formal models, employing them to derive correct executives by formal computation and synthesis, integrated in a compilation process. We therefore use models throughout the design flow from specification to implementation, in particular by automatic generation of embeddable executives.

3.2.2. Applicative Needs

Applicative needs initially come from the fields of safety-critical systems (*e.g.*, avionics, nuclear) and complex systems (telecommunications), embedded in an environment with which they strongly interact (comprising aspects of computer science and control theory). Fields with less criticality, or which support variable degrees of quality of service, such as in the multi-media domain, can also take advantage of methodologies that improve the quality and reliability of software, and reduce the costs of test and correction in the design.

Industrial acceptance, the dissemination, and the deployment of the formal techniques inevitably depend on the usability of such techniques by specialists in the application domain — and not in formal techniques themselves — and also on the integration in the whole design process, which concerns very different problems and techniques. Application domains where the actors are ready to employ specialists in formal methods or advanced control theory are still uncommon. Even then, design methods based on the systematic application of these theoretical results are not ripe. In fields like industrial control, where the use of PLC (Programmable Logic Controller [29]) is dominant, this question can be decisive.

Essential elements in this direction are the proposal of realistic formal models, validated by experiments, of the usual entities in control theory, and functionalities (*i.e.*, algorithms) that correspond indeed to services useful for the designer. Take, for example, the compilation and optimization taking into account the platforms of execution, the possible failures, or the interactions between the defined automatic control and its implementation. In these areas, there are functionalities that commercial tools do not have yet, and to which our results contribute.

[6] http://www.synalp.org
[7] http://www-verimag.imag.fr/SYNCHRONE
[8] http://www.inria.fr/equipes/aoste
[9] http://www.irisa.fr/espresso/Polychrony
[10] http://www.di.ens.fr/~pouzet/lucid-synchrone/
[11] http://www-rocq.inria.fr/syndex

3.2.3. Our Approach

We are proposing effective trade-offs between, on the one hand, expressiveness and formal power, and on the other hand, usability and automation. We focus on the area of specification and construction of correct real-time executives for discrete and continuous control, while keeping an interest in tackling major open problems, relating to the deployment of formal techniques in computer science, especially at the border with control theory. Regarding the applications, we propose new automated functionalities, to be provided to the users in integrated design and programming environments.

3.3. Main Research Directions

The overall consistency of our approach comes from the fact that the main research directions address, under different aspects, the specification and generation of safe real-time control executives based on *formal models*.

We explore this field by linking, on the one hand, the techniques we use, with on the other hand, the functionalities we want to offer. We are interested in questions related to:

Component-Based Design. We investigate two main directions: (i) compositional analysis and design techniques; (ii) adapter synthesis and converter verification.

Programming for embedded systems. Programming for embedded real-time systems is considered within POP ART along three axes: (i) synchronous programming languages, (ii) aspect-oriented programming, (iii) static analysis (type systems, abstract interpretation, ...).

Dependable embedded systems. Here we address the following research axes: (i) static multiprocessor scheduling for fault-tolerance, (ii) multi-criteria scheduling for reliability, (iii) automatic program transformations, (iv) formal methods for fault-tolerant real-time systems.

The creation of easily usable models aims at giving the user the role rather of a pilot than of a mechanics *i.e.*, to offer her/him pre-defined functionalities which respond to concrete demands, for example in the generation of fault tolerant or distributed executives, by the intermediary use of dedicated environments and languages.

The proposal of validated models with respect to their faithful representation of the application domain is done through case studies in collaboration with our partners, where the typical multidisciplinarity of questions across control theory and computer science is exploited.

3.3.1. Component-Based Design

Component-based construction techniques are crucial to overcome the complexity of embedded systems design. However, two major obstacles need to be addressed: the heterogeneous nature of the models, and the lack of results to guarantee correction of the composed system.

The heterogeneity of embedded systems comes from the need to integrate components using different models of computation, communication, and execution, at different levels of abstraction and different time scales. The BIP component framework [5] has been designed, in cooperation with VERIMAG, to support this heterogeneous nature of embedded systems.

Our work focuses on the underlying analysis and construction algorithms, in particular compositional techniques and approaches ensuring correctness by construction (adapter synthesis, strategy mapping). This work is motivated by the strong need for formal, heterogeneous component frameworks in embedded systems design.

3.3.2. Programming for Embedded Systems

Programming for embedded real-time systems is considered along three directions: (i) synchronous programming languages to implement real-time systems; (ii) aspect-oriented programming to specify non-functional properties separately from the base program; (iii) abstract interpretation to ensure safety properties of programs at compile time. We advocate the need for well defined programming languages to design embedded real-time systems with correct-by-construction guarantees, such as bounded time and bounded memory execution. Our original contribution resides in programming languages inheriting features from both synchronous languages

and functional languages. We contribute to the compiler of the HEPTAGON language (whose main inventor is Marc Pouzet, ENS Paris, PARKAS team), the key features of which are: data-flow formal synchronous semantics, strong typing, modular compilation. In particular, we are working on type systems for the clock calculus and the spatial modular distribution.

The goal of Aspect-Oriented Programming (AOP) is to isolate aspects (such as security, synchronization, or error handling) that cross-cut the program basic functionality and whose implementation usually yields tangled code. In AOP, such aspects are specified *separately* and integrated into the program by an automatic transformation process called *weaving*. Although this paradigm has great practical potential, it still lacks formalization, and undisciplined uses make reasoning on programs very difficult. Our work on AOP addresses these issues by studying foundational issues of AOP (semantics, analysis, verification) and by considering domain-specific aspects (availability or fault tolerance aspects) as formal properties.

Finally, the aim of the verification activity in POP ART is to check safety properties on programs, with emphasis on the analysis of the values of data variables (numerical variables, memory heap), mainly in the context of embedded and control-command systems that exibit concurrency features. The applications are not only the proof of functional properties on programs, but also test selection and generation, program transformation, controller synthesis, and fault-tolerance. Our approach is based on abstract interpretation, which consists in inferring properties of the program by solving semantic equations on abstract domains. Much effort is spent on implementing developed techniques in tools for experimentation and diffusion.

3.3.3. Dependable Embedded Systems

Embedded systems must often satisfy safety critical constraints. We address this issue by providing methods and algorithms to design embedded real-time systems with guarantees on their fault-tolerance and/or reliability level.

A first research direction concerns static multiprocessor scheduling of an application specification on a distributed target architecture. We increase the fault-tolerance level of the system by replicating the computations and the communications, and we schedule the redundant computations according to the faults to be tolerated. We also optimize the schedule *w.r.t.* several criteria, including the schedule length, the reliability, and the power consumption.

A second research direction concerns the fault-tolerance management, by reconfigurating the system (for instance by migrating the tasks that were running on a processor upon the failure of this processor) following objectives of fault-tolerance, consistent execution, functionality fulfillment, boundedness and optimality of response time. We base such formal methods on discrete controller synthesis.

A third research direction concerns AOP to weave fault-tolerance aspects in programs and electronic circuits (seen as synthesizable HDL programs) as mentioned in the previous section. A first step in this direction has been the design of automatic transformation method for fault tolerance, which implement a limited (but nonetheless interesting) form of AOP.

PROSECCO Project-Team

3. Scientific Foundations

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 (in 2003, 2008, 2009, and 2011) 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 [52]; it has motivated a serious research effort on the formal analysis of cryptographic protocols, starting with [50] 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* [46]. 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 [44], [38], correspondences (including authentication) [45], and observational equivalences [43]. 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 [13], JFK [39], and Web Services Security [42]. 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 [48], results which were extended by formal analysis work using a Dolev-Yao style model of the standard [49]. 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

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# [13]. 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 [40] 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 [41]. F* [53] 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.

An alternative approach is to directly verify cryptographic applications in the computational model by typing. A recent work [51] 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.

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. 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 F7 to verify their correctness.

S4 Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

The research work of the team is built on top of solid foundations, mainly, algebraic, combinatorial or logical theories of transition systems. These theories cover several sorts of systems which have been studied during the last thirty years: sequential, concurrent, synchronous or asynchronous. They aim at modeling the behavior of finite or infinite systems (usually by abstracting computations on data), with a particular focus on the control flow which rules state changes in these systems. Systems can be autonomous or reactive, that is, embedded in an environment with which the system interacts, both receiving an input flow, and emitting an output flow of events and data. System specifications can be explicit (for instance, when the system is specified by an automaton, extensively defined by a set of states and a set of transitions), or implicit (symbolic transition rules, usually parameterized by state or control variables; partially-synchronized products of finite transition systems; Petri nets; systems of equations constraining the transitions of synchronous reactive systems, according to their input flows; etc.). Specifications can be non-ambiguous, meaning that they fully define at most one system (this holds in the previous cases), or they can be ambiguous, in which case more than one system is conforming to the specification (for instance, when the system is described by logical formulas in the modal mu-calculus, or when the system is described by a set of scenario diagrams, such as *Sequence Diagrams* or *Message Sequence Charts*).

Systems can be described in two ways: either the state structure is described, or only the behavior is described. Both descriptions are often possible (this is the case for formal languages, automata, products of automata, or Petri nets), and moving from one representation to the other is achieved by folding/unfolding operations.

Another taxonomy criteria is the concurrency these models can encompass. Automata usually describe sequential systems. Concurrency in synchronous systems is usually not considered. In contrast, Petri nets or partially-synchronized products of automata are concurrent. When these models are transformed, concurrency can be either preserved, reflected or even, infused. An interesting case is whenever the target architecture requires distributing events among several processes. There, communication-efficient implementations require that concurrency is preserved as far as possible and that, at the same time, causality relations are also preserved. These notions of causality and independence are best studied in models such as concurrent automata, Petri nets or Mazurkiewicz trace languages.

Here are our sources of inspiration regarding formal mathematical tools:

1. Jan van Leeuwen (ed.), *Handbook of Theoretical Computer Science - Volume B: Formal Models and Semantics*, Elsevier, 1990.
2. Jörg Desel, Wolfgang Reisig and Grzegorz Rozenberg (eds.), *Lectures on Concurrency and Petri nets*, Lecture Notes in Computer Science, Vol. 3098, Springer, 2004.
3. Volker Diekert and Grzegorz Rozenberg (eds.), *The Book of Traces*, World Scientific, 1995.
4. André Arnold and Damian Niwinski, *Rudiments of Mu-Calculus*, North-Holland, 2001.
5. Gérard Berry, *Synchronous languages for hardware and software reactive systems - Hardware Description Languages and their Applications*, Chapman and Hall, 1997.

Our research exploits decidability or undecidability results on these models (for instance, inclusion of regular languages, bisimilarity on automata, reachability on Petri nets, validity of a formula in the mu-calculus, etc.) and also, representation theorems which provide effective translations from one model to another. For instance, Zielonka's theorem yields an algorithm which maps regular trace languages to partially-synchronized products of finite automata. Another example is the theory of regions, which provides methods for mapping finite or infinite automata, languages, or even *High-Level Message Sequence Charts* to Petri nets. A further example concerns the mu-calculus, in which algorithms computing winning strategies for parity games can be used to synthesize supervisory control of discrete event systems.

Our research aims at providing effective representation theorems, with a particular emphasis on algorithms and tools which, given an instance of one model, synthesize an instance of another model. In particular we have contributed a theory, several algorithms and a tool for synthesizing Petri nets from finite or infinite automata, regular languages, or languages of *High-Level Message Sequence Charts*. This also applies to our work on supervisory control of discrete event systems. In this framework, the problem is to compute a system (the controller) such that its partially-synchronized product with a given system (the plant) satisfies a given behavioral property (control objective, such as a regular language or satisfaction of a mu-calculus formula).

Software engineers often face problems similar to *service adaptation* or *component interfacing*, which in turn, often reduce to particular instances of system synthesis or supervisory control problems.

SECRET Project-Team

3. Scientific Foundations

3.1. Scientific foundations

Our research work is mainly devoted to the design and analysis of cryptographic algorithms. 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... Most of our work mix 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.

SECSI Project-Team

3. Scientific Foundations

3.1. Foundations

Computer security has become more and more pressing as a concern since the mid 1990s. There are several reasons to this: cryptography is no longer a *chasse réservée* of the military, and has become ubiquitous; and computer networks (e.g., the Internet) have grown considerably and have generated numerous opportunities for attacks and misbehaviors, notably.

The aim of the SECSI project is to *develop logic-based verification techniques for security properties of computer systems and networks*. Let us explain what this means, and what this does not mean.

First, the scope of the research at SECSI started as a rather broad subset of computer security, although the core of SECSI's activities has always been on verifying cryptographic protocols.

We took this for granted in 2006, and decided to concentrate on the latter. This already includes a vast number of concerns.

First, there is a plethora of distinct *security properties* one may wish to verify. Beyond the standard properties of secrecy (weak or strong forms), or authentication, one considers anonymity, fairness in contract-signing, and the subtle security properties involved in electronic voting such as accountability, receipt-freeness, resistance to coercion, or user verifiability. Some of these properties are trace properties, some are not, and are therefore more complex to state and verify.

Second, there are many available *models*. SECSI started with the rather simple symbolic models of security known today as Dolev-Yao models. One must then look at process algebra models (spi-calculus, applied pi-calculus), which allow for a symbolic treatment of more complex properties, especially those that are not trace properties. And one must also look at the computational models favored by cryptographers, e.g., the game-based approaches and the universal composability/simulatability approaches. They are more realistic in terms of security, but less directly amenable to automated verification. One of the features of computational models that makes them more complex is the need for computing, and bounding probabilities of certain events. This led us into contributing to the field of verification of probabilistic systems. One must also look at the relations between these models.

Third, there are many important *applications*. While SECSI started looking at the rather simple and now mundane confidentiality and authentication protocols, two important application domains have emerged: the verification of electronic voting protocols, and the verification of cryptographic APIs.

Apart from cryptographic protocols, the initial vision of the SECSI project was that computer security, being a global concern, should be taken as a whole, as far as possible. This is why one of the initial objectives of SECSI included topic in intrusion detection, again seen from the logical point of view.

One should remember the following. First, one of the key phrases in the SECSI motto is "logic-based". It is a founding theme of SECSI that logic matters in security, and opportunities are to be grabbed. Another key phrase is "verification techniques". The expertise of SECSI is not in designing protocols or security architectures. Verifying protocols, formally, is an arduous task already, and has proved to be an extremely rich area.

3.2. Objectives

SECSI has five objectives:

- Objective 1: symbolic verification of cryptographic protocols. Tree-automata based methods, automated deduction, and approximate/exact cryptographic protocol verification in the Dolev-Yao model. Enriching the Dolev-Yao model with algebraic theories, and associated decision problems.

- Objective 2: verification of cryptographic protocols in computational models. Computational soundness of formal models (Dolev-Yao, applied pi-calculus).
- Objective 3: security of group protocols, fair exchange, voting and other protocols. Other security properties, other security models. Security properties based on notions of indistinguishability.
- Objective 4: probabilistic transition systems. Security in the presence of probabilistic and demonic non-deterministic choices.
- Objective 5: intrusion detection, network and host protection in the large.

TASC Project-Team

3. Scientific Foundations

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://www.emn.fr/x-info/sdemasse/gccat/index.html). 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 350 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.

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 combinate the expressive power of global constraints with local search frameworks.

3.2.4. Solvers

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 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). 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.). Since 2009 year, due to the hiring of G. 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). This work around the design and the development of solvers thus forms the fourth direction of basic research of the project.

TOCCATA Team

3. Scientific Foundations

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, a new axis of research that we propose is to develop certified tools: 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 [5], including both interactive proving and automated solving [9]. 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 [62][4]. This is why certification of numerical programs is another axis of this proposition.

The *Toccata* project emphasizes two new axes of research: certified tools and verification of numerical programs. Additionally we want to 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. Certified Programs,
2. Certified Tools,
3. Certified Numerical Programs.

The reader should be aware that the word “Certified” in this scientific programme means “verified by a formal specification and a formal proof that the program meets this specification”. This differs from the standard meaning of “Certified” in an industrial context which is that it conforms to a rigorous process and/or a norm.

3.2. Certified Programs

This theme of research builds upon our expertise on the development of methods and tools for proving programs, from source codes annotated with specifications to proofs. In the past years, 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 certified 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 certified 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 Certified 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 needs also the design 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 certified 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. [90]). The well-known proposals for solving it include *Separation logic* [107], *implicit dynamic frames* [105], and *considerate reasoning* [106]. 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 [108] or a system of theory cloning in our system *Why3* [1]; 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 [69], [88], [50].

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 generate *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 want to 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 [6].

A key challenge is also in the better 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 [42], [32]. 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 certified components will require higher-order specifications. There are a few solutions that were proposed yet, that amount to encode higher-order goals in first-order ones [88].

Generally speaking, there are several theories, interesting for program verification, that we would like to add as built-in decision procedures in a SMT context. First, although there already exist decision procedures for variants of bit-vectors, these 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 [12]. 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 [68]. We now have a prototype for 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 would like to follow is to design 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* [99] 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

In this theme concerning certified 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 in 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 [45], to cryptographic programs [49]. We plan to use proof techniques on applications related to databases. We also have plans to support low-level programs such as assembly code [81], [102] 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 [1] [24]) or abstract interpretation techniques (new project Cafein that will start in 2013) and also for discharging proof obligations coming from other systems like Atelier B [29] (new project BWare, Section 8.2.1).

3.3. Certified Tools

The goal of this theme is to guarantee the soundness of the tools we develop. Indeed 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 analysers 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 [74] formalizing term rewriting systems, the Alea library [45] for the formalization of randomized algorithms, several libraries formalizing floating-point numbers (Floats [58], Gappalib [97], and now Flocq [5] which unifies the formers). Second we recently conducted the development of a certified decision procedure [93] that corresponds to a core part of *Alt-Ergo*, and a certified verification condition generator for a language [28] similar to *Why*. On-going work aims at building, still in *Coq*, a certified VC generator for C annotated in ACSL [55], based on the operational semantics formalized in the CompCert certified compiler project [92].

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 more freedom. For example, it should allow one to use a bit of side-effects when the underlying algorithm justify 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 certified libraries of programs, but also libraries of specifications (e.g. theories of sets, maps, etc.).

[1] http://cubicle.lri.fr/

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.

A first attempt in considering such an issue was done in [29]. 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 will be an important topic of the new 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* [93] 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. Currently, *Gappa* can be asked to produce a *Coq* proof of its given goal, so as to check *a posteriori* its soundness. Indeed, the idea of producing a trace is not contradictory with certifying the tool. For very complex decision procedures, the goal of developing a certified proof search might be too ambitious, and the production of an internal trace is a general technique that might be used as a workaround: it suffices to instrument the proof search and to develop a certified trace checker to be used by the tool before it gives an answer. We used this approach in the past for certified proofs of termination of rewriting systems [75]. This is also a technique that is used internally in CompCert for some passes of compilation [92].

3.3.4. Certified VC generation

The other kind of tools that we would like to certify are the VC generators. This will be a continuation of the on-going work on developing in *Coq* a certified VC generator for C code annotated in ACSL. We would like to develop such a generator in *Why3* instead of *Coq*. As before, this will build 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* [91], and a generator for unstructured programs like assembly, that would operate on an arbitrary control-flow graph.

On a longer term, it would be interesting 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 [70] 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 are not first-class objects in *Coq*. 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 [52].

3.4. Certified 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 [65], [109], [61], [104], [64], [59], [98], [96] but also the web galleries of certified programs at our Web page [2], the Hisseo project [3], S. Boldo's page [4], and industrial case studies in the U3CAT ANR project. The ability to express such complex properties comes from models developed in *Coq* [5]. 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 [4]. However, the proofs can get difficult even on short programs, and to achieve them some automation is needed, and obtained by combining SMT solvers and *Gappa* [62], [79], [46][9]. 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 [66], [102].

The directions of research concerning floating-point programs that we want to 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 on-going development of the Flocq library should be continued 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 smart algorithms to compute efficiently with floats, which operate at the bit-level. The formalization of real numbers need to be 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 tend to be confused by floating-point arithmetic [31]. 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

The goal is to handle floating-point programs that are related to continuous quantities. This includes numerical analysis programs we have already worked on [14] [61][3]. 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 this kind of programs. 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 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 than four years, 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

[2] http://toccata.lri.fr/gallery/index.en.html
[3] http://hisseo.saclay.inria.fr/
[4] http://www.lri.fr/~sboldo/research.html

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 refsec:atp, directly as a certified program in *Coq* or *Why3*.

TRIO Project-Team

3. Scientific Foundations

3.1. Fondation

In order to check for the timing behavior and the reliability of distributed systems, the TRIO team developed several techniques based on deterministic approaches ; in particular, we apply and extend analytical evaluation of worst case response times and when necessary, e.g. for large-scale communication systems as Internet based applications, we use techniques based on network calculus.

When the environment might lead to hazards (e.g. electromagnetic interferences causing transmission errors and bit-flips in memory), or when some characteristics of the system are not perfectly known or foreseeable beforehand, we model and analyze the uncertainties using stochastic models, for instance, models of the frame transmission patterns or models of the transmission errors. In the context of real time computing, we are in general much more interested by worst-case results over a given time window than by average and asymptotic results, and dedicated analyses in that area have been developed in our team over the last 10 years.

In the design of discrete event systems with hard real time constraints, the scheduling of the system's activities is of crucial importance. This means that we have to devise scheduling policies that ensure the respect of time constraints on line and / or optimize the behavior of the system according to some other application-dependent performance criteria.

In order to foster the best quality for programs, their understanding has to be automated, or at made significantly easier. Thus, we focus on analyzing and modeling program code, program structure and program behavior, and presenting these pieces of information to the user (in our case, program designers and program developers). Modeling user interaction is to come as well.

In the design of embedded, autonomous systems, power and energy usage is of paramount importance. We thus strive to model energy usage, based on actual hardware, and derive context-aware optimizations to decrease peak power and overall energy usage.

TYPICAL Project-Team

3. Scientific Foundations

3.1. Logical formalisms

A proof system implements a logical formalism in the way a compiler implements a programming language. Similarly, the choice of the formalism is crucial for the success of the proof system. One of the main line of research of the team is to study or invent type theories that are well-adapted to the formalization of mathematics. For instance a crucial property of a proof system is its correctness, hence the importance of the study of the models of the meta-theory of the Coq proof assistant. An other issue is the interoperability of the various proof systems used to formalize mathematics in the world-wide community of users of proof assistants, and the design of a system which could serve as a back-end to front-end implementing various formalisms and proof languages.

3.2. Libraries of formalized mathematics

It is well known that advanced mathematics can play a crucial role in the design and correctness of sophisticated and sometimes critical software. In some cases, using a proof system is the only option to mechanize the correctness of such programs; this can require the formalization of a wide variety of mathematical theories, and a careful design of these formal libraries for them to be maintainable, combinable and reusable. Furthermore, the ability to formalize advanced contemporary mathematics is still a form of ultimate quality tests for proof systems, and also a way to gain visibility. One of our objectives is to make modern and large pieces of mathematics available as usable formal libraries. Recent examples of complex proofs (Four Color Theorem, Kepler conjecture, classification of finite groups, Fermat theorem) challenge the way the mathematical literature is refereed and published. We think that the development of these formal libraries of mathematics may also change the way certain mathematical result become accepted as theorems. Crafting large bodies of formalized mathematics is a challenging task. These libraries obey similar requirements as software : modularity and usability stem from appropriate data-structures, design patterns and corpus of lemmas. But the appropriate methodology leading to the relevant solutions is often far from obvious, and this is where research has to be done and know-how has to be gained. Up to recently, formal developments were seldom collaborative and rarely benefitted from reusable previous work. The maturity of proof assistants is now sufficient to envision a more modern conception of formal software, as required by large scale verification projects like T. Hales' proof of the Kepler conjecture or the Feit-Thompson theorem. Several members of the TypiCal team are committed in such big formalization projects, or in more specific but related side projects.

3.3. Proof search and automated decision procedures

Interactive proof assistants provide a very expressive logical formalism, rich enough to allow extremely precise descriptions of complex objects like the meta theory of a programming language, a model of C compiler, or the proof of the Four Color Theorem. This description includes logical statements of the properties required by the objects of interest but also their formal proofs, checked by the merciless proof-checker of the system, which should be a small hence trusted piece of code. These systems provide the highest formal guarantee, for instance, of the correctness with respect to the mathematical specification of a code.

Proof-search is a central issue in such a formalization of mathematics. It is also a common aspect of automated reasoning and high-level programming paradigms such as Logic programming. However specific applications commonly involve specific logics or theories, like for instance linear arithmetic. Whether or not such a logical framework can express these at all, it is unlikely that its generic proof-search mechanisms can replace the methods that are specific to a logic or theory. Either because this specific domain lies outside the reach of generic proof-search or simply because generic proof-search is less efficient therein than a purpose-made procedure (typically a decision procedure).

But to enlarge the scope where a specific method applies, one can combine both generic proof search mechanisms with specific methods. We hence investigate how to craft formal proof producing decision procedures in the context of an interactive proof assistant. This activity includes understanding the impact of proof-search mechanism (polarization, focusing, etc.), the implementation of efficient connections between domain specific automated decision procedures (SMT solvers, polynomial optimization tools, etc.) with a proof assistant, and the combination of these two aspects in the design a unique logical framework where a generic notion of proof-search could serve each of the above purposes.

VEGAS Project-Team (section vide)

VERIDIS Project-Team

3. Scientific Foundations

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 for the proved development of concurrent and distributed systems and algorithms. Our common objective is to advance the state of the art of combining interactive with automated methods resulting in powerful tools for the (semi-)automatic verification of distributed systems and protocols. Our techniques and tools will support methods for the formal development of trustworthy distributed systems that are grounded in mathematically precise semantics and that scale to algorithms relevant for practical applications.

The VeriDis members from Saarbrücken are developing Spass [7], one of the leading automated theorem provers for first-order logic based on the superposition calculus [33]. Recent extensions to the system include the integration of dedicated reasoning procedures for specific theories, such as linear arithmetic [44], [31], 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 [45] and automated reasoning mechanisms these induce.

The VeriDis members from Nancy develop veriT [1], an SMT (Satisfiability Modulo Theories [35]) 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$^+$ [41] specifications. Our prover relies on a declarative proof language and includes several automatic backends [3].

3.2. Methodology of proved system development

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], [6], and in applying them to concrete use cases. In particular, the concept of *refinement* [30], [34], [43] in state-based modeling formalisms is central to our approach. Its basic idea is to derive an algorithm or implementation by providing 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 reduce the number of generated proof obligations and/or support their proof 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, potentially again with support of SMT procedures for specific sub-problems, or with the help of interactive proof guidance.

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 current system paradigms, including ad-hoc and overlay networks or peer-to-peer systems.

VERTECS Project-Team

3. Scientific Foundations

3.1. Underlying models

The formal models we use are mainly automata-like structures such as labelled transition systems (LTS) and some of their extensions: an LTS is a tuple $M = (Q, \Lambda, \rightarrow, q_o)$ where Q is a non-empty set of states; $q_o \in Q$ is the initial state; A is the alphabet of actions, $\rightarrow \subseteq Q \times \Lambda \times Q$ is the transition relation. These models are adapted for testing and controller synthesis.

To model reactive systems in the testing context, we use Input/Output labeled transition systems (IOLTS for short). In this setting, the interactions between the system and its environment (where the tester lies) must be partitioned into inputs (controlled by the environment), outputs (observed by the environment), and internal (non observable) events modeling the internal behavior of the system. The alphabet Λ is then partitioned into $\Lambda_! \cup \Lambda_? \cup \mathcal{T}$ where $\Lambda_!$ is the alphabet of outputs, $\Lambda_?$ the alphabet of inputs, and $\mathcal{T}$ the alphabet of internal actions.

In the controller synthesis theory, we also distinguish between controllable and uncontrollable events ($\Lambda = \Lambda_c \cup \Lambda_{uc}$), observable and unobservable events ($\Lambda = \Lambda_O \cup \mathcal{T}$).

In the context of verification, we also use Timed Automata. A timed automaton is a tuple $A = (L, X, E, \mathcal{I})$ where L is a set of locations, X is a set of clocks whose valuations are positive real numbers, $E \subseteq L \times \mathcal{G}(X) \times 2^X \times L$ is a finite set of edges composed of a source and a target state, a guard given by a finite conjunction of expressions of the form $x \sim c$ where x is a clock, c is a natural number and $\sim \in \{<, \leq, =, \geq, >\}$, a set of resetting clocks, and $\mathcal{I} : \mathcal{L} \rightarrow \mathcal{G}(\mathcal{X})$ assigns an invariant to each location [22]. The semantics of a timed automaton is given by a (infinite states) labelled transition system whose states are composed of a location and a valuation of clocks.

Also, for verification purposes, we use graph grammars that are a general tool to define families of graphs. Such grammars are formed by a set of rules whose left-hand sides are hyperedges and right-hand sides are hypergraphs. For graphs with finite degree, these grammars characterise transition graphs of pushdown automata (the correspondence between graphs generated by grammars and transition graphs of pushdown automata is bijective). Graph grammars provide a simple yet powerful setting to define and study infinite state systems.

In order to cope with models closer to practical specification languages, we also need higher level models encompassing both control and data aspects. We defined (input-output) symbolic transition systems ((IO)STS), which are extensions of (IO)LTS that convey or operate on data (i.e., program variables, communication parameters, symbolic constants) through message passing, guards, and assignments. Formally, an IOSTS is a tuple (V, Θ, Σ, T), where V is a set of variables (including a counter variable encoding the control structure), Θ is the initial condition defined by a predicate on V, Σ is the finite alphabet of actions, where each action has a signature (just like in IOLTS, Σ can be partitioned as e.g. $\Sigma_? \cup \Sigma_! \cup \Sigma_\tau$), T is a finite set of symbolic transitions of the form $t = (a, p, G, A)$ where a is an action (possibly with a polarity reflecting its input/output/internal nature), p is a tuple of communication parameters, G is a guard defined by a predicate on p and V, and A is an assignment of variables. The semantics of IOSTS is defined in terms of (IO)LTS where states are vectors of values of variables, and transitions between them are labelled with instantiated actions (action with valued communication parameter). This (IO)LTS semantics allows us to perform syntactical transformations at the (IO)STS level while ensuring semantical properties at the (IO)LTS level. We also consider extensions of these models with added features such as recursion, fifo channels, etc. An alternative to IOSTS for specifying systems with data variables is the model of synchronous dataflow equations.

Our research is based on well established theories: conformance testing, supervisory control, abstract interpretation, and theorem proving. Most of the algorithms that we employ take their origins in these theories:

- graph traversal algorithms (breadth first, depth first, strongly connected components, ...). We use these algorithms for verification as well as test generation and control synthesis.
- BDDs (Binary Decision Diagrams) algorithms, for manipulating Boolean formulae, and their MTBDDs (Multi-Terminal Decision Diagrams) extension for manipulating more general functions. We use these algorithms for verification, test generation and control.
- abstract interpretation algorithms, specifically in the abstract domain of convex polyhedra (for example, Chernikova's algorithm for the computation of dual forms). Such algorithms are used in verification and test generation.
- logical decision algorithms, such as satisfiability of formulas in Presburger arithmetics. We use these algorithms during generation and execution of symbolic test cases.

3.2. Verification

Verification in its full generality consists in checking that a system, which is specified by a formal model, satisfies a required property. Verification takes place in our research in two ways: on the one hand, a large part of our work, and in particular controller synthesis and conformance testing, relies on the ability to solve some verification problems. Many of these problems reduce to reachability and coreachability questions on a formal model (a state s is *reachable from an initial state* s_i if an execution starting from s_i can lead to s; s is *coreachable from a final state* s_f if an execution starting from s can lead to s_f). These are important cases of verification problems, as they correspond to the verification of safety properties.

On the other hand we investigate verification on its own in the context of complex systems. For expressivity purposes, it is necessary to be able to describe faithfully and to deal with complex systems. Some particular aspects require the use of infinite state models. For example asynchronous communications with unknown transfer delay (and thus arbitrary large number of messages in transit) are correctly modeled by unbounded FIFO queues, and real time systems require the use of continuous variables which evolve with time. Apart from these aspects requiring infinite state data structure, systems often include uncertain or random behaviours (such as failures, actions from the environment), which it make sense to model through probabilities. To encompass these aspects, we are interested in the verification of systems equipped with infinite data structures and/or probabilistic features.

When the state space of the system is infinite, or when we try to evaluate performances, standard model-checking techniques (essentially graph algorithms) are not sufficient. For large or infinite state spaces, symbolic model-checking or approximation techniques are used. Symbolic verification is based on efficient representations of sets of states and permits exact model-checking of some well-formed infinite-state systems. However, for feasibility reasons, it is often mandatory to use approximate computations, either by computing a finite abstraction and resort to graph algorithms, or preferably by using more sophisticated abstract interpretation techniques. For systems with stochastic aspects, a quantitative analysis has to be performed, in order to evaluate the performances. Here again, either symbolic techniques (e.g. by grouping states with similar behaviour) or approximation techniques should be used.

We detail below verification topics we are interested in: abstract interpretation, quantitative model-checking and analysis of systems defined by graph grammars.

3.2.1. Abstract interpretation and data handling

Most problems in test generation or controller synthesis reduce to state reachability and state coreachability problems which can be solved by fixpoint computations of the form $x = F(x), x \in C$ where C is a lattice. In the case of reachability analysis, if we denote by S the state space of the considered program, C is the lattice $\wp(S)$ of sets of states, ordered by inclusion, and F is roughly the "*successor states*" function defined by the program.

The big change induced by taking into account the data and not only the (finite) control of the systems under study is that the fixpoints become uncomputable. The undecidability is overcome by resorting to approximations, using the theoretical framework of Abstract Interpretation [24]. The fundamental principles of Abstract Interpretation are:

1. to substitute to the *concrete domain*C a simpler *abstract domain*A (static approximation) and to transpose the fixpoint equation into the abstract domain, so that one has to solve an equation $y = G(y), y \in A$;
2. to use a *widening operator* (dynamic approximation) to make the iterative computation of the least fixpoint of G converge after a finite number of steps to some upper-approximation (more precisely, a post-fixpoint).

Approximations are conservative so that the obtained result is an upper-approximation of the exact result. In simple cases the state space that should be abstracted has a simple structure, but this may be more complicated when variables belong to different data types (Booleans, numerics, arrays) and when it is necessary to establish *relations* between the values of different types.

3.2.2. Model-checking quantitative systems

Model-checking techniques for finite-state systems are now quite developed, and a current challenge is to adapt them as much as possible to infinite-state systems. We detail below two types of models we are interested in: timed automata and infinite-state probabilistic systems.

Model-checking timed automata The model of timed automata, introduced by Alur and Dill in the 90's [22] is commonly used to represent real-time systems. Timed automata consist of an extension of finite automata with continuous variables, called clocks, that evolve synchronously with time, and can be tested and reset along an execution. Despite their uncountable state space, checking reachability, and more generally ω-regular properties, is decidable *via* the construction of a finite abstraction, the so-called region automaton. The recent developments in model-checking timed automata have aimed at modelling and verifying quantitative aspects encompassing timing constraints, for example costs, probabilities, frequencies. These quantitative questions demand advanced techniques that go far beyond the classical methods.

Model-checking infinite state probabilistic systems Model-checking techniques for finite state probabilistic systems are now quite developed. Given a finite state Markov chain, for example, one can check whether some property holds almost surely (i.e. the set of executions violating the property is negligible), and one can even compute (or at leat approximate as close as wanted) the probability that some property holds. In general, these techniques cannot be adapted to infinite state probabilistic systems, just as model-checking algorithms for finite state systems do not carry over to infinite state systems. For systems exhibiting complex data structures (such as unbounded queues, continuous clocks) and uncertainty modeled by probabilities, it can thus be hard to design model-checking algorithms. However, in some cases, especially when considering qualitative verification, symbolic methods can lead to exact results. Qualitative questions aim neither at computing nore at approximating a probability, but are only concerned with almost-sure or non neglectible behaviours (that is events of probability either one or non zero). In some cases, qualitative model-checking can be derived from a combination of techniques for infinite state systems (such as abstractions) with methods for finite state probabilistic systems. However, when one is interested in computing (or rather approximating) precise probability values (neither 0 nor 1), exact methods are scarce. To deal with these questions, we either try to restrict to classes of systems where exact computations can be made, or look for approximation algorithms.

3.2.3. Analysis of infinite state systems defined by graph grammars

Currently, many techniques (reachability, model checking, ...) from finite state systems have been generalised to pushdown systems, that can be modeled by graph grammars. Several such extensions heavily depend on the actual definition of the pushdown automata, for example, how many top stack symbols may be read, or whether the existence of ε-transitions (silent transitions) is allowed. Many of these restrictions do not affect the actual structure of the graph, and interesting properties like reachability or satisfiability (of a formula) only depend on the structure of a graph.

Deterministic graph grammars enable us to focus on structural properties of systems. The connection with finite graph algorithms is often straightforward: for example reachability is obtained by iterating the finite graph algorithm iterated on the right hand sides of the rules. On the other hand, extending these grammars with time or probabilities is not straightforward: qualitative values associated to different copies (in the graph) of the same vertex (in the grammar) may differ, introducing complex equations. Furthermore, the fact that the left-hand sides of rules are single hyperarcs is a strong restriction. But removing this restriction would lead to non-recursive graphs. Identifying decidable families of graphs defined by contextual graph grammars is also very challenging.

3.3. Automatic test generation

We are mainly interested in conformance testing, which consists in checking whether a black box implementation under test (the real system that is only known by its interface) behaves correctly with respect to its specification (the reference which specifies the intended behavior of the system). In the line of model-based testing, we use formal specifications and their underlying models to unambiguously define the intended behavior of the system, to formally define conformance and to design test case generation algorithms. The difficult problems are to generate test cases that correctly identify faults (the oracle problem) and, as exhaustiveness is impossible to reach in practice, to select an adequate subset of test cases that are likely to detect faults. Hereafter we detail some elements of the models, theories and algorithms we use.
We use IOLTS (or IOSTS) as formal models for specifications, implementations, test purposes, and test cases. We adapt a well established theory of conformance testing [30], which formally defines conformance as a relation between formal models of specifications and implementations. This conformance relation, called **ioco** compares the visible behaviors (called *suspension traces*) of the implementation I (denoted by $STraces(I)$) with those of the specification S ($STraces(S)$). Suspension traces are sequence of inputs, outputs or quiescence (absence of action denoted by δ), thus abstracting away internal behaviors that cannot be observed by testers. Intuitively, I $ioco$ S if after a suspension trace of the specification, the implementation I can only show outputs and quiescences of the specification S. We re-formulated ioco as a partial inclusion of visible behaviors as follows:

$$I \; ioco \; S \Leftrightarrow STraces(I) \cap [STraces(S).\Lambda_!^\delta \setminus STraces(S)] = \varnothing.$$

In other words, suspension traces of I which are suspension traces of S prolongated by an output or quiescence, should still be suspension traces of S.

Interestingly, this characterization presents conformance with respect to S as a safety property of suspension traces of I. The negation of this property is charaterized by a *canonical tester*$Can(S)$ which recognizes exactly $[STraces(S).\Lambda_!^\delta \setminus STraces(S)]$, the set of non-conformant suspension traces. This *canonical tester* also serves as a basis for test selection.

Test cases are processes executed against implementations in order to detect non-conformance. They are also formalized by IOLTS (or IOSTS) with special states indicating *verdicts*. The execution of test cases against implementations is formalized by a parallel composition with synchronization on common actions. A *Fail* verdict means that the implementation under test (IUT) is rejected and should correspond to non-conformance, a *Pass* verdict means that the IUT exhibited a correct behavior and some specific targeted behaviour has been observed, while an *Inconclusive* verdict is given to a correct behavior that is not targeted.

Test suites (sets of test cases) are required to exhibit some properties relating the verdict they produce to the conformance relation. Soundness means that only non conformant implementations should be rejected by a test suite and exhaustiveness means that every non conformant implementation may be rejected by the test suite. Soundness is not difficult to obtain, but exhaustiveness is not possible in practice and one has to select test cases.

Test selection is often based on the coverage of some criteria (state coverage, transition coverage, etc). But test cases are often associated with *test purposes* describing some abstract behaviors targeted by a test case. In our framework, test purposes are specified as IOLTS (or IOSTS) associated with marked states or dedicated variables, giving them the status of automata or observers accepting runs (or sequences of actions or suspension traces). Selection of test cases amounts to selecting traces of the canonical tester accepted by the test purpose. The resulting test case is then both an observer of the negation of a safety property (non-conformance wrt. S), and an observer of a reachability property (acceptance by the test purpose). Selection can be reduced to a model-checking problem where one wants to identify states (and transitions between them) which are both reachable from the initial state and co-reachable from the accepting states. We have proved that these algorithms ensure soundness. Moreover the (infinite) set of all possibly generated test cases is also exhaustive. Apart from these theoretical results, our algorithms are designed to be as efficient as possible in order to be able to scale up to real applications.

Our first test generation algorithms are based on enumerative techniques, thus adapted to IOLTS models, and optimized to fight the state-space explosion problem. On-the-fly algorithms where designed and implemented in the TGV tool, which consist in computing co-reachable states from a target state during a lazy exploration of the set of reachable states in a product of the specification and the test purpose [25]. However, this enumerative technique suffers from some limitations when specification models contain data.

More recently, we have explored symbolic test generation techniques for IOSTS specifications [29]. The objective is to avoid the state space explosion problem induced by the enumeration of values of variables and communication parameters. The idea consists in computing a test case under the form of an *IOSTS*, i.e., a reactive program in which the operations on data are kept in a symbolic form. Test selection is still based on test purposes (also described as IOSTS) and involves syntactical transformations of IOSTS models that should ensure properties of their IOLTS semantics. However, most of the operations involved in test generation (determinisation, reachability, and coreachability) become undecidable. For determinisation we employ heuristics that allow us to solve the so-called bounded observable non-determinism (i.e., the result of an internal choice can be detected after finitely many observable actions). The product is defined syntactically. Finally test selection is performed as a syntactical transformation of transitions which is based on a semantical reachability and co-reachability analysis. As both problems are undecidable for IOSTS, syntactical transformations are guided by over-approximations using abstract interpretation techniques. Nevertheless, these over-approximations still ensure soundness of test cases [26]. These techniques are implemented in the STG tool (see 5.1), with an interface with NBAC used for abstract interpretation.

3.4. Control synthesis

The supervisory control problem is concerned with ensuring (not only checking) that a computer-operated system works correctly. More precisely, given a system model and a required property, the problem is to control the model's behavior, by coupling it to a supervisor, such that the controlled system satisfies the property [28]. The models used are LTSs and the associated languages, where one makes a distinction between *controllable* and *non-controllable* actions and between *observable* and *non-observable* actions. Typically, the controlled system is constrained by the supervisor, which can block on the system's controllable actions in order to force it to behave as specified by the property. The control synthesis problem can be seen as a constructive verification problem: building a supervisor that prevents the system from violating a property. Several kinds of properties can be enforced such as reachability, invariance (i.e. safety), attractivity, etc. Techniques adapted from model checking are used to compute the supervisor. Optimality must be taken into account as one often wants to obtain a supervisor that constrains the system as few as possible.

Supervisory control theory overview. Supervisory control theory deals with control of Discrete Event Systems. In this theory, the behavior of the system S is assumed not to be fully satisfactory. Hence, it has to be reduced by means of a feedback control (named Supervisor or Controller) in order to achieve a given set of requirements [28]. Namely, if S denotes the model of the system and Φ a safety property to be enforced on S, the problem consists of computing a supervisor $\mathcal{C}$ such that

$$S \| \mathcal{C} \models \Phi \tag{1}$$

where $\|$ is the classical parallel composition of LTSs. Given S, some events of S are said to be uncontrollable (Σ_{uc}), i.e., the occurrence of these events cannot be prevented by a supervisor, while the others are controllable (Σ_c). It means that all the supervisors satisfying (1) are not good candidates. The behavior of the controlled system must respect an additional condition that happens to be similar to the $ioco$ conformance relation previously defined in 3.3 . This condition is called the *controllability condition* and it may be stated as

$$\mathcal{L}(S \| \mathcal{C}) \Sigma_{uc} \cap \mathcal{L}(S) \subseteq \mathcal{L}(S \| \mathcal{C}) \tag{2}$$

Namely, when acting on S, a supervisor is not allowed to disable uncontrollable events. Given a safety property Φ, that can be modeled by an LTS A_Φ, there actually exist many different supervisors satisyfing both (1) and (2). Among all the valid supervisors, we are interested in computing the supremal one, ie the one that restricts the system as few as possible. It has been shown in [28] that such a supervisor always exists and is unique. It gives access to a behavior of the controlled system that is called the supremal controllable sub-language of A_Φ w.r.t. S and Σ_{uc}. In some situations, it may also be interesting to force the controlled system to be non-blocking (See [28] for details).

The underlying techniques are similar to the ones used for Automatic Test Generation. They consist of computing the product of the system model and A_Φ and to remove the states of the product that may lead to subsequent states violating the property by triggering only uncontrollable events.

ALEA Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

This idea of analyzing nature systems and transferring the underlying principles into stochastic algorithms and technical implementations is one of the central component of the ALEA team project. Adapting nature mechanisms and biological capabilities clearly provides a better understanding of the real processes, and it also improves the performance and the power of engineers devices. Our project is centered on both the understanding of biological processes in terms of mathematical, physical and chemical models, and on the other hand, on the use of these biology inspired stochastic algorithms to solve complex engineering problems.

There is a huge series of virtual interfaces, robotic devices, numerical schemes and stochastic algorithms which were invented mimicking biological processes or simulating natural mechanisms. The terminology *"mimicking or simulating"* doesn't really mean to find an exact copy of natural processes, but *to elaborate the mathematical principles so that they can be abstracted from the original biological or physical model.* In our context, the whole series of evolutionary type principles discussed in previous sections can be abstracted into only three different and natural classes of stochastic algorithms, depending on the nature of the biology-inspired interaction mechanism used in the stochastic evolution model. These three stochastic search models are listed below :

1) *Branching and interacting particle systems (birth and death chains, spatial branching processes, mean-field interaction between generations):*

The first generation of adaptive branching-selection algorithms is very often built on the same genetic type paradigm: When exploring a state space with many particles, we duplicate better fitted individuals at the expense of light particles with poor fitness die. From a computational point of view, we generate a large number of random problem solvers. Each one is then rated according to a fitness or performance function defined by the developer. Mimicking natural selection, an evolutionary algorithm selects the best solvers in each generation and breeds them.

2) *Reinforced random walks and self-interacting chains (reinforced learning strategies, interaction processes with respect to the occupation measure of the past visited sites):*

This type of reinforcement is observed frequently in nature and society, where "beneficial" interactions with the past history tend to be repeated. A new class of historical mean field type interpretation models of reinforced processes were developed by the team project leader in a pair of articles [56], [55]. Self interaction gives the opportunity to build new stochastic search algorithms with the ability to, in a sense, re-initialized their exploration from the past, re-starting from some better fitted initial value already met in the past [57], [58].

3) *Random tree based stochastic exploration models (coalescent and genealogical tree search explorations techniques on path space):*

The last generation of stochastic random tree models is concerned with biology-inspired algorithms on paths and excursions spaces. These genealogical adaptive search algorithms coincide with genetic type particle models in excursion spaces. They have been applied with success in generating the excursion distributions of Markov processes evolving in critical and rare event regimes, as well as in path estimation and related smoothing problems arising in advanced signal processing (cf. [53] and references therein). We underline the fact that the complete mathematical analysis of these random tree models, including their long time behavior, their propagations of chaos properties, as well as their combinatorial structures are far from being completed. This class of genealogical tree based models has been introduced in [54] for solving smoothing problems and more generally Feynman-Kac semigroups on path spaces, see also [52], [53], and references therein.

APICS Project-Team

3. Scientific Foundations

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 behaviour. Perhaps the most basic example of such a problem is harmonic identification of a stable linear dynamical system: the transfer-function f is holomorphic in the right half-pane, which means it satisfies there the Cauchy-Riemann equation $\overline{\partial} f = 0$, and in principle f can be recovered from its values on the imaginary axis, *e.g.* by 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 [69]. 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, it satisfies $\overline{\partial} f = \sum_1^n a_j \delta_{z_j}$ where the z_j are its poles, and finding the domain of holomorphy (*i.e.* locating the z_j) amounts to solve a (degenerate) free-boundary inverse problem, this time on the left half-plane. To address these questions, the team has developed a two-step approach as follows.

Step 1: To determine a complete model, that is, one which is defined for every frequency, in a sufficiently flexible function class (*e.g.* Hardy spaces). This ill-posed issue requires regularization, for instance constraints on the behaviour 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. Derivation of the complete model is important to achieve stability of the reduced one.

Step 1 makes connection with 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 output, see section 3.3.2.2 . It also makes contact with the topology of rational functions, to count critical points and to derive bounds, see section 3.3.2 . Moreover, this step raises issues in approximation theory regarding the rate of convergence and whether the singularities of the approximant (*i.e.* its poles) converge to the singularities of the approximated function; this is where logarithmic potential theory becomes effective, see section 3.3.3 .

Iterating the previous steps coupled with a sensitivity analysis yields a tuning procedure which was first demonstrated in [77] on resonant microwave filters.

Similar steps can be taken to approach design problems in frequency domain, replacing measured behaviour by desired behaviour. However, describing achievable responses from the design parameters at hand is generally cumbersome, and most constructive techniques rely on rather specific criteria adapted to the physics of the problem. This is especailly true of circuits and filters, whose design classically appeals to standard polynomial extremal problems and realization procedures from system theory [70], [55]. APICS is active in this field, where we introduced the use of Zolotarev-like problems for microwave multiband filter design. We currently favor interpolation techniques because of their transparency with respect to parameter use, see section 3.2.2 .

In another connection, the 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; portion of the boundary where data are not available may be unknown, in which case we meet a free boundary problem. This framework for 2-D non-destructive control was first advocated in [59] and subsequently received considerable attention. This framework makes it 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 [1] [79]. All these questions are particular instances of the

so-called inverse potential problem, where a measure μ has to be recovered from 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. In any case, by construction, the operator applied to the potential yields back the measure.

Inverse potential problems are severely indeterminate because infinitely many measures within an open set produce the same field outside this set [68]. In step 1 above we implicitly removed this indeterminacy by requiring that the measure be supported on the boundary (because we seek a function holomorphic throughout the right half space), and in step 2 by requiring, say, in case of rational approximation that the measure be discrete in the left half-plane. The same discreteness assumption prevails in 3-D inverse source problems. 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 this 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, the team initiated the use of steps 1 and 2 above, along with singularity analysis, to approach issues of nondestructive control in 2 and 3-D [41] [6], [2]. We are currently engaged in two kinds of generalization, further described in section 3.2.1 . The first one deals with non-constant conductivities, where Cauchy-Riemann equations for holomorphic functions are replaced by conjugate Beltrami equations for pseudo-holomorphic functions; there we seek applications to plasma confinement. The other one lies with inverse source problems for Laplace's equation in 3-D, where holomorphic functions are replaced by harmonic gradients, developing applications to EEG/MEG and inverse magnetization problems in paleomagnetism, see section 4.2

The main approximation-theoretic tools developed by APICS to get to grips with issues mentioned so far are outlined in section 3.3 . In section 3.2 to come, we make more precise which problems are considered and for which applications.

3.2. Range of inverse problems

3.2.1. Elliptic partial differential equations (PDE)

Participants: Laurent Baratchart, Slah Chaabi, Juliette Leblond, Ana-Maria Nicu, Dmitry Ponomarev, Elodie Pozzi.

This work is done in collaboration with Alexander Borichev (Univ. Provence).

Reconstructing Dirichlet-Neumann boundary conditions for a function harmonic in a plane domain when these are known on a strict 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 determined by their values on boundary subsets of positive linear measure, which is the framework for problem (P) in section 3.3.1 . Such problems naturally arise in nondestructive testing of 2-D (or cylindical) materials from partial electrical measurements on the boundary. Indeed, the ratio between tangential and normal currents (so-called Robin coefficient) tells about corrosion of the material. Solving problem (P) where ψ is chosen to be the response of some uncorroded piece with identical shape allows one to approach such questions, and this was an initial application of holomorphic extremal problems to non-destructive control [56], [52].

[1] 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

A recent application by the team deals with non-constant conductivity over a doubly connected domain, E being the outer boundary. Measuring Dirichlet-Neumann data on E, we wanted to check whether the solution is constant on the inner boundary. We first had to define and study Hardy spaces of the conjugate Beltrami equation, of which the conductivity equation is the compatibility condition (just like Laplace's equation is the compatibility condition of the Cauchy-Riemann system). This was done in references [5] and [35]. Then, solving an obvious modification of problem (P) allows one to numerically check what we want. Further, the value of this extremal problem defines a criterion on inner boundaries, and subsequently a descent algorithm was set up to improve the initial boundary into one where the solution is closer to being constant, thereby trying to solve a free boundary problem..

When the domain is regarded as separating the edge of a tokamak's vessel from the plasma (rotational symmetry makes this a 2-D problem), the procedure just described suits plasma control from magnetic confinement. It was successfully applied in collaboration with CEA (the French nuclear agency) and the University of Nice (JAD Lab.) to data from *Tore Supra* [58], see section 6.2 . This procedure is fast because no numerical integration of the underlying PDE is needed, as an explicit basis of solutions to the conjugate Beltrami equation was found in this case.

Three-dimensional versions of step 1 in section 3.1 are also considered, namely 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. Such questions arise naturally in connection with neurosciences and medical imaging (electroencephalography, EEG) or in paleomagnetism (analysis of rocks magnetization) [2] [37], see section 6.1 . They are not yet as developed as the 2-D case where the power of complex analysis is at work, but considerable progress was made over the last years through methods of harmonic analysis and operator theory.

The team is also concerned with non-destructive control problems of localizing defaults such as cracks, sources or occlusions in a planar or 3-dimensional domain, from boundary data (which may correspond to thermal, electrical, or magnetic measurements). These defaults can be expressed as a lack of analyticity of the solution of the associated Dirichlet-Neumann problem and we approach them using techniques of best rational or meromorphic approximation on the boundary of the object [4] [16], see sections 3.3.2 and 4.2 . In fact, the way singularities of the approximant relate to the singularities of the approximated function is an all-pervasive theme in approximation theory, and for appropriate classes of functions the location of the poles of a best rational approximant can be used as an estimator of the singularities of the approximated function (see section 6.1). This circle of ideas is much in the spirit of step 2 in section 3.1 .

A genuine 3-dimensional theory of approximation by discrete potentials, though, is still in its infancy.

3.2.2. Systems, transfer and scattering

Participants: Laurent Baratchart, Sylvain Chevillard, Sanda Lefteriu, Martine Olivi, Fabien Seyfert.

Through initial contacts with CNES, the French space agency, the team came to work on identification-for-tuning of microwave electromagnetic filters used in space telecommunications (see section 4.3). The problem was to recover, from band-limited frequency measurements, the physical parameters of the device under examination. The latter consists of interconnected dual-mode resonant cavities with negligible loss, hence its scattering matrix is modelled 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 enters the scene, through the so-called *realization* process mapping a rational transfer function in the frequency domain to a state-space representation of the underlying system as a 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, and in particular the Hilbert space H^2, provide a framework to transform this classical ill-posed issue into a series of well-posed analytic and meromorphic approximation problems. The procedure sketched in section 3.1 now 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 in the Hardy space H^2 of the right half-plane a problem analogous to (P) in section 3.3.1 , taking into account prior knowledge on the decay of the response outside the bandwidth, see [18] for details.
2. From this stable model, a rational stable approximation of appropriate degree is computed. For this a descent method is used on the relatively compact manifold of inner matrices of given size and degree, using a novel parametrization of stable transfer functions [18].
3. From this rational model, realizations meeting certain constraints imposed by the technology in use are computed (see section 6.3). These constraints typically come from the nature and topology of the equivalent electrical network used to model the filter. This network is composed of resonators, coupled to each other by some specific coupling topology. Performing this realization step for given coupling topology can be recast, under appropriate compatibility conditions [8], as the problem of solving a zero-dimensional multivariate polynomial system. To tackle this problem in practice, we use Groebner basis techniques as well as continuation methods as implemented in the Dedale-HF software (5.4).

Let us also mention that extensions of classical coupling matrix theory to frequency-dependent (reactive) couplings have lately been carried-out [1] for wide-band design applications, but further study is needed to make them effective.

Subsequently APICS started investigating issues pertaining to filter design rather than identification. Given the topology of the filter, a basic problem is to find the optimal response with respect to amplitude specifications in frequency domain bearing on rejection, transmission and group delay of scattering parameters. Generalizing the approach based on Tchebychev polynomials for single band filters, we recast the problem of multi-band response synthesis in terms of a generalization of classical Zolotarev min-max problem [30] to rational functions [11]. Thanks to quasi-convexity, the latter can be solved efficiently using iterative methods relying on linear programming. These are implemented in the software easy-FF (see section 5.5).

Later, investigations by the team extended to design and identification of more complex microwave devices, like multiplexers and routers, which connect several filters through wave guides. Schur analysis plays an important role in such studies, which is no surprise since 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 [76], who devised a recursive test to check for contractivity of a holomorphic function in the disk. Generalizations thereof turned out to be very efficient to parametrize solutions to contractive interpolation problems subject to a well-known compatibility condition (positive definiteness of the so-called Pick matrix) [32]. Schur analysis became quite popular in electrical engineering, as the Schur recursion precisely describes how to chain two-port circuits.

Dwelling on this, members of the team contributed to differential parametrizations (atlases of charts) of loss-less matrix functions to the theory [31][12], [10]. They are of fundamental use in our rational approximation software RARL2 (see section 5.1). Schur analysis is also instrumental to approach de-embedding issues considered in section 6.4 , and provides further background to current studies by the team of synthesis and adaptation problems for multiplexers. At the heart of the latter lies a variant of contractive interpolation with degree constraint introduced in [62].

We also mention the role played by multipoint Schur analysis in the team's investigation of spectral representation for certain non-stationary discrete stochastic processes [3], [36].

Recently, in collaboration with UPV (Bilbao), our attention was driven by CNES, to questions of stability relative to high-frequency amplifiers, see section 7.2 . Contrary to previously mentioned devices, these are *active* components. The amplifier can be linearized at a functioning point and admittances of the corresponding electrical network can be computed at various frequencies, using the so-called harmonic balance method. The

goal is to check for stability of this linearised model. The latter is composed of lumped electrical elements namely inductors, capacitors, negative *and* positive reactors, transmission lines, and commanded current sources. Research so far focused on determining the algebraic structure of admittance functions, and setting up a function-theoretic framework to analyse them. In particular, much effort was put on realistic assumptions under which a stable/unstable decomposition can be claimed in $H^2 \oplus \overline{H^2}$ (see section 6.5). Under them, the unstable part of the elements under examination is rational and we expect to bring valuable estimates of stability to the designer using the general scheme in section 3.1 .

3.3. Approximation of boundary data

Participants: Laurent Baratchart, Sylvain Chevillard, Juliette Leblond, Martine Olivi, Dmitry Ponomarev, Elodie Pozzi, Fabien Seyfert.

The following people are collaborating with us on these topics: Bernard Hanzon (Univ. Cork, Ireland), Jean-Paul Marmorat (Centre de mathématiques appliquées (CMA), École des Mines de Paris), Jonathan Partington (Univ. Leeds, UK), Ralf Peeters (Univ. Maastricht, NL), Edward Saff (Vanderbilt University, Nashville, USA), Herbert Stahl (TFH Berlin), Maxim Yattselev (Univ. Oregon at Eugene, USA).

3.3.1. Best constrained 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 $D \subset \mathbb{R}^2$, we want to recover a holomorphic function from its values on a subset of the boundary of D. Using conformal mapping, it is convenient for the discussion to normalize D. So, in the simply connected case, we fix D to be the unit disk with boundary the unit circle T. We denote by H^p the Hardy space of exponent p which is the closure of polynomials in the L^p-norm on the circle 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 in D approximately matching measured values f 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 behaviour capturing *a priori* assumptions on the behaviour of the model off K, while M is some admissible deviation from them. The value of p reflects the type of stability which is sought and how much one wants to smoothen the data. The choice of L^p classes is well-adapted to handling pointwise measurements.

To fix terminology we refer to (P) as a *bounded extremal problem*. As shown in [40], [42], [47], for $1 < p \leq \infty$, the solution to this convex infinite-dimensional optimization problem can be obtained upon iterating with respect to a Lagrange parameter the solution to spectral equations for some appropriate Hankel and Toeplitz operators. These equations in turn involve the solution to the standard extremal problem below best approximation problem (P_0) below [61]:

> (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 $p = 1$ of (P) is essentially open.

Various modifications of (P) have been studied in order to meet specific needs. For instance when dealing with loss-less transfer functions (see section 4.3), one may want to express the constraint on $T \setminus K$ in a pointwise manner: $|g - \psi| \leq M$ a.e. on $T \setminus K$, see [43]. In this form, it comes close to (but still is different from) H^∞ frequency optimization methods for control [64], [75].

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 recovering 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). For $p = 2$ the solution is analysed in [66]. 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 flux. Then, the Lagrange parameter indicates which deformation should be applied on the inner contour in order to improve data fitting.

This is 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 non constant. In this case Hardy spaces in problem (P) become those of a so-called conjugate or real Beltrami equation [65], which are studied for $1 < p < \infty$ in [5], [35]. Expansions of solutions needed to constructively handle such issues have been carried out in [58].

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 a $\mathbb{R}^n$-valued vector V field 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 convenient substitute of holomorphic Hardy spaces is provided by Stein-Weiss Hardy spaces of harmonic gradients [78]. Conformal maps are no longer available in $\mathbb{R}^n$ for $n > 2$ and other geometries have not been much studied so far. On the ball, the analog of problem (P) is

> $(P1)$ 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, and $M > 0$; 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 $p = 2$, spherical harmonics offer a reasonable substitute to Fourier expansions and problem $(P1)$ was solved in [2], together with its natural analog on a shell. The solution generalizes the Toeplitz operator approach to bounded extremal problems [40], and constructive aspects of the procedure (harmonic 3-D projection, Kelvin and Riesz transformation, spherical harmonics) were derived. An important ingredient is a refinement of the 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 $H(B)$, a vector field in $H^p(\mathbb{R}^n \setminus \overline{B}$, and a tangential divergence free vector field. If $p = 1$ or $p = \infty$, L^p must be replaced respectively by the real Hardy space H^1 and the bounded mean oscillation space BMO, and H^∞ should be modified accordingly. This decomposition was fully discussed in [37] (for the case of the half-space) where it plays a fundamental role.

Problem $(P1)$ is still under investigation in the case $p = \infty$, where even the case where $O = S$ is pending because a substitute of the Adamjan-Arov-Krein theory [72] is still to be built in dimension greater than 2.

Such problems arise in connection with source recovery in electro/mgneto encephalography and paleomagnetism, as discussed in sections 3.2.1 and 4.2 .

3.3.2. Best meromorphic and rational approximation

The techniques explained in this section are used to solve step 2 in section 3.2*via* conformal mapping and subsequently instrumental to approach inverse boundary value problems for Poisson equation $\Delta u = \mu$, where μ is some (unknown) distribution.

3.3.2.1. Scalar meromorphic and rational approximation

Let as before D designate the unit disk, and T the unit circle. We further put R_N for the set of rational functions with at most N poles in D, which allows us to define the meromorphic functions in $L^p(T)$ as the 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 continuous f it is known how to solve (P_N) in closed form. The unique solution is given by AAK theory (named after Adamjan, Arov and Krein), that connects the spectral decomposition of Hankel operators with best approximation in Hankel norm [72]. This theory allows one to express g_N in terms of the singular vectors of the Hankel operator with symbol f. The continuity of g_N as a function of f only holds for stronger norms than uniform.

The case $p = 2$ is of special importance. In particular when $f \in \overline{H}^2$, the Hardy space of exponent 2 of the *complement* of D in the complex plane (by definition, $h(z)$ belongs to $\overline{H}^p$ if, and only if $h(1/z)$ belongs to H^p), then (P_N) reduces to rational approximation. Moreover, it turns out that the associated solution $g_N \in R_N$ has no pole outside D, hence it is a *stable* rational approximant to f. However, in contrast with the situation when $p = \infty$, this approximant may *not* be unique.

The former Miaou project (predecessor of Apics) has designed an adapted 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. Roughly speaking, it is a gradient algorithm that proceeds recursively with respect to the order N of the approximant, in a compact region of the parameter space [34]. Although it has proved effective in all applications carried out so far (see sections 4.2 , 4.3), it is not known 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 undertook a long-term study of the number and nature of critical points, in which tools from differential topology and operator theory team up with classical approximation theory. The main discovery is that the nature of the critical points (*e.g.*, local minima, saddles...) depends on the decrease of the interpolation error to f as N increases [44]. Based on this, sufficient conditions have been developed for a local minimum to be unique. These conditions are hard to use in practice because they require strong estimates of the approximation error. These are often difficult to obtain for a given function, and are usually only valid for large N. Examples where uniqueness or asymptotic uniqueness has been proved this way include transfer functions of relaxation systems (*i.e.* Markov functions) [48] and more generally Cauchy integrals over hyperbolic geodesic arcs [50] and certain entire functions [46].

An analog to AAK theory has been carried out for $2 \leq p < \infty$ [47]. Although not computationally as powerful, it can be used to derive lower bounds and helps analysing the behaviour of poles. When $1 \leq p < 2$, problem (P_N) is still fairly open.

A common feature to all these problems is that critical point equations express non-Hermitian orthogonality relations for the denominator of the approximant. This makes connection with interpolation theory [51][7] and is used in an essential manner to assess the behaviour of the poles of the approximants to functions with branchpoint-type 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 the approximation of a vector field with gradients of potentials generated by N point masses instead of meromorphic functions. The issue is by no means understood at present, and is a major endeavour of future research problems.

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. Such questions have become over years an increasingly significant part of the team's activity (see section 4.3). For instance, convergence properties of multipoint Schur approximants, which are rational interpolants preserving contractivity of a function, were analysed in [3]. Such approximants are useful in prediction theory of stochastic processes, but since they interpolate inside the domain of holomorphy they are of limited use in frequency design.

In another connection, the generalization to several arcs of classical Zolotarev problems [74] is an achievement by the team which is useful for multiband synthesis [11]. Still, though the modulus of the response is the first concern in filter design, variation of the phase must nevertheless remain under control to avoid unacceptable distortion of the signal. This specific but important issue has less structure and was approached

using constrained optimization; a dedicated code has been developed under contract with the CNES (see section 5.5).

3.3.2.2. Matrix-valued rational approximation

Matrix-valued approximation is necessary for handling systems with several inputs and outputs, and it generates substantial additional difficulties with respect to scalar approximation, theoretically as well as algorithmically. In the matrix case, the McMillan degree (*i.e.* the degree of a minimal realization in the System-Theoretic sense) generalizes the degree.

The problem we want to consider reads: *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 [34], mentioned in section 3.3.2.1 , generalizes to the matrix-valued situation [60]. The first difficulty here consists in the parametrization of transfer matrices of given McMillan degree n, and the inner matrices (*i.e.* matrix-valued functions that are analytic in the unit disk and unitary on the circle) of degree n. The latter enter the picture in an essential manner as they play the role of the denominator in a fractional representation of transfer matrices (using the so-called Douglas-Shapiro-Shields factorization). The set of inner matrices of given degree has the structure of 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 (parametrization valid in a neighborhood of a point), and we must handle changes of charts in the course of the algorithm. Such parametrization can be obtained from interpolation theory and Schur type algorithms, the parameters being interpolation vectors or matrices ([31], [10], [12]). Some of them are particularly interesting to compute realizations and achieve filter synthesis ([10] [12]). Rational approximation software codes have been developed in the team (see sections 5.1).

Difficulties relative to multiple local minima naturally 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 sought degree or small perturbations thereof (the consistency problem) was solved in [45]. The case of matrix-valued Markov functions, the first example beyond rational functions, was treated in [33].

Let us stress that the algorithms mentioned above are first to handle rational approximation in the matrix case in a way that converges to local minima, while meeting stability constraints on the approximant.

3.3.3. Behavior of poles of meromorphic approximants

Participant: Laurent Baratchart.

The following people collaborate with us on this subject: Herbert Stahl (TFH Berlin), Maxim Yattselev (Univ. Oregon at Eugene, USA).

We refer here to the behaviour 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. If one normalizes the contour to be the unit circle T, we are back to the framework of section 3.3.2.1 and to problem (P_N); invariance of the problem under conformal mapping was established in [6]. Research so far has focused on functions whose singular set inside the contour is zero or one-dimensional.

Generally speaking, the behaviour of poles is particularly important in meromorphic approximation to obtain error rates as the degree goes large and to tackle constructive issues like uniqueness. As explained in section 3.2.1 , we consider this issue in connection with approximation of the solution to a Dirichlet-Neumann problem, so as to extract information on the singularities. 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 initialization of more precise but heavier numerical optimizations.

As regards crack detection or source recovery, the approach in question boils down to analysing the behaviour 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 ([6], [38], [14]) that the poles of the approximants accumulate on some extremal contour of minimum weighted energy linkings the singular points of the crack, or the sources [41]. Moreover, the asymptotic density of the poles turns out to be the Green equilibrium distribution of this contour in D, hence puts heavy charge around the singular points (in particular at the endpoints) which are therefore well localized if one is able to approximate in sufficiently high degree (this is where the method could fail).

The case of two-dimensional singularities is still an outstanding open problem.

It is interesting that inverse source problems inside a sphere or an ellipsoid in 3-D can be attacked with the above 2-D techniques, as applied to planar sections (see section 6.1).

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 around 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 Solya is not a long-term objective of APICS.

ASPI Project-Team

3. Scientific Foundations

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) [50], 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* [52], *Monte Carlo filter* [57] or *condensation* (conditional density propagation) algorithm [54], 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 [53], [47], visual tracking [54], mobile robotics [48], [71], ubiquitous computing and ambient intelligence, sensor networks, risk evaluation and simulation of rare events [51], genetics, molecular simulation [49], etc.

Other examples of the many applications of particle filtering can be found in the contributed volume [34] 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 [35] can be found, and in the textbook [67] 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 [66]. A recent overview can also be found in [37].

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 [42], [40], 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 [74]), see e.g. the survey article [42] or the recent textbook [40], and references therein

> convergence in L^p, convergence as empirical processes indexed by classes of functions, uniform convergence in time, see also [63], [64], central limit theorem, see also [59], 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 [38], 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 [43] 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 [41], [64], 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 [58]. 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 [55],
- 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 [56]: 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 [65] that (i) the parameter dependent probability distributions of the observations are locally asymptotically normal (LAN) [61], 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 , 5.6 , 5.10 and 5.11 .

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 [60], 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 [68], [39], 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 [69] 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. Scientific Foundations

3.1. Numerical schemes for fluid mechanics

Participants: Rémi Abgrall, Mario Ricchiuto, Pietro Marco Congedo, Cécile Dobrzynzki, Héloïse Beaugendre, Pierre-Henri Maire, Luc Mieussens.

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 that will be discussed later. Another application is the influence of fluid flow on noise production. Problems in aeroacoustics are indeed becoming more and more important in everyday life. In some occasions, one needs specific numerical tools to take into account *e.g.* a fluids' exotic equation of state, or because the amount of required computational resources becomes huge, as in unsteady flows. 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, a situation where commercial tools can only provide a very crude answer.

It is a fact that there are many commercial codes. They allow users to simulate a lot of different flow types. The quality of the results is however far from optimal in many cases. Moreover, the numerical technology implemented in these codes is often 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.*

In order to simulate efficiently these complex physical problems, we are working on some fundamental aspects of the numerical analysis of non linear hyperbolic problems. *Our goal is to develop more accurate and more efficient schemes that can adapt to modern computer architectures*.

More precisely, *we are working on 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. An effort for developing a more predictive tool for multiphase compressible flows is also underway. Within this project, several advancements have been performed, i.e. considering a more complete systems of equations including viscosity, working on the

thermodynamic modeling of complex fluids, and developing stochastic methods for uncertainty quantification in compressible flows.

We expect within a few years 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, unsteady aerodynamics and aeroacoustics associated to both external and internal compressible flows, compressible ideal and non-ideal MHD (in relation with the ITER project), and the behavior 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.

Our research in numerical algorithms has led to the development of the RealfluiDS platform which is described in section 5.3 . New software developments are under way in the field of free surface flows and complex materials modeling. These developments are performed in the code SLOWS (Shallow-water fLOWS) for free surface flows, and in the solver COCA (CodeOxydationCompositesAutocicatrisants) for the simulation of the self-healing process in composite materials. These developments will be described in sections 5.10 and 5.2 .

This work is supported by the EU-Strep IDIHOM, various research contracts and in part by the ANEMOS project and the ANR-Emergence RealFluids grant. A large part of the team also beneficiates of the ADDECCO ERC grant.

3.2. Uncertainty quantification

Participants: Rémi Abgrall, Mario Ricchiuto, Pietro Marco Congedo.

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 developped 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 or semi-intrusive methods, trying to define an unified framework for obtained a reliable and accurate numerical solution at a moderate computational cost. Dealing with high dimensional representation of stochastic inputs in design optimiza- tion is 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.

This part of our activities is supported by the ERC grant ADDECCO, the ANR-MN project UFO and the associated team AQUARIUS.

3.3. Meshes and scalable discrete data structures

Participants: Cécile Dobrzynski, Sébastien Fourestier, Algiane Froehly, Cédric Lachat, François Pellegrini.

3.3.1. Adaptive dynamic mesh 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.

Performing such re-meshing in parallel creates 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 [51] 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 [52], [50], 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 boules that can be processed concurrently, each by a sequential remesher. It requires to devise scalable algorithms for building such boules, scheduling them on the available processors, reconstructing the remeshed mesh and redistributing its data. This research started within the context of the PhD of Cédric Lachat, funded by a CORDI grant of EPI `PUMAS` and is continued thanks to a funding by ADT grant `El Gaucho`.

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 5.9) 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 research takes place within the context of the PhD of Sébastien Fourestier.

BIPOP Project-Team

3. Scientific Foundations

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 [4], 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 [10]) 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], [11].
- 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 [13].

CAD Team

3. Scientific Foundations

3.1. Geometry continuity and ε Geometry Continuity

The mathematical background of parametric surfaces is Differential Geometry. In differential geometry, Riemann (1826 1866), Shiing-Shen Chern (1911 2004), continuities play a very important kernel role. In 1980s, more and more engineering design using geometry modeling softwares found the problems of the parametric continuities. And the order of the parametric continuity depends on how the curve is parameterized. To day, engineers and scientists try to find a kind of continuities, which are the intuitive intrinsic properties of curves and surfaces, and the orders of the continuities are independent of the parameterization.

G-Continuity could be defined as the smoothness properties of a curve or a surface that are more than its order of differentiability. This problem is complex and progress in this domain is very slow. We proposed new ways to make through the bottleneck. Furthermore, we also wanted to fill the gap between the traditional mathematics and modern computer science. Hence, we developed the theories of epsilon-geometry continuities to accommodate the representation and the rounding errors of float-point arithmetic, and design new geometric modeling operators under the constraints of epsilon-geometry continuities.

3.2. Two main challenges in Computer Aided Design

3.2.1. Robustness tolerance, error control

Based on this theoretical contribution, we also proposed several elegant solutions to the most important challenges in Computer Aided Design (see Lees A Piegl. "Ten challenges in Computer-Aided-Design". Jal of CAD 2005. 37 (4): 461-470): robustness, tolerances, error control, geometric arrangement, beautification and modelling of complex shapes. During CAD processes one uses a myriad of tolerances, many of which are not directly related to the actual manufacturing process. Some interesting questions here include: What are the most relevant machining tolerances? How to set the army of computational tolerances, e.g. those of systems of equations, to guarantee machining within the required accuracy? How tolerances in different spaces, e.g. in model space and in parameter space, are related. Numerical instabilities also account for the majority of computational errors in commercial CAD systems.

The problems related to robustness haunt every programmer who has ever worked on commercial systems. Fixing numerical bugs can be very frustrating, and often times results in patching up the code simply because no solution exists to remedy the problem.

3.2.2. Geometry beautification, Geometry operators and Shape generation

Although geometric uncertainties are related to robustness and tolerance, there are a number of extra issues well worth deeper investigations. Geometric arrangements are full of special cases. The most notable ones are: cases of touch, overlapping, containment, etc.; cases of parallelism, perpendicularity, coincidence, etc.; axes of symmetrical data, data clustering, dense or sparse data, etc.; cases of degeneracy, discontinuity, inconsistencies, etc.; problems with cracks, excess material, lack of detail, etc. In just about any code that deals with geometry, the number of special cases is significantly larger than the general ones. Data explosion is the result of careless selection of the methods, e.g. parameter space-based sampling, and improper implementation, e.g. recursive algorithms. Some of the relevant issues are: sampling: over sampling, sampling in incorrect places, etc; procedural definitions, e.g. lofting a large set of curves or merging surfaces may result in an explosion of control points.

Furthermore, although CAD processes are supposed to produce valid and "made to order" models, the reality is that most (if not all) models are rough and require post-processing, i.e. beautification. Some of the most frequently needed tasks are: removing unwanted edges, corners, cracks, etc.; removing bumps, oscillations, curvature extremes, etc.; healing incorrect models, e.g. removing holes in triangulations; smoothing, fairing, re-shaping, etc.

3.3. Computer Graphics

In Computer Graphics, objectives were to prove the capability of the team to address some topics as Computational Photography, Rendering and Computer Animation. Work in Progress in these topics are described in the following chapter.

CAGIRE Team

3. Scientific Foundations

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 $Re^{3/4}$. The smallest scales may have a certain effect on the largest ones which implies that an accurate framework for the computation of flows must take into account all these scales. This can be achieved either by solving directly the Navier-Stokes equations (Direct numerical simulations or DNS) or by first applying a time filtering (Reynolds Average Navier-Stokes or RANS) or a spatial filtering operator to the Navier-Stokes equations (large-eddy simulations or LES). The new terms brought about by the filtering operator have to be modelled. 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. By simulating the large scale structures while modelling the smallest ones supposed to be more isotropic, LES 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). Along these lines, our objective is to develop a DNS tool to simulate a jet in crossflow configuration which is the generic flow of an aeronautical combustion chamber as far as its effusion cooling is concerned.

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 [11]. 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 [31] and first studied by Lesaint and Raviart [24]. 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 [15], [14], one dimensional systems [13], multidimensional scalar conservation laws [12], and multidimensional systems [16]. For the same system, we can also cite the work of [18], [22], 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 [10], 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 [9]. 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 [20]. The symmetric nature of the discretization is particularly well suited with mesh adaptation by means of the adjoint equation resolution [21]. Last, we note that the discontinuous Galerkin method was already successfully tested in [17] at Direct Numerical Simulation scale for very moderate Reynolds, and also by Munz'team in Stuttgart [25], 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 for example: efficient shock capturing term methods for supersonic flows, high order discretization of curved boundaries, or low Mach behaviour of these schemes (this last point will be detailed in the next subsection). Another drawback of the discontinuous Galerkin methods is that they are very computationally costly, due to the accurate representation of the solution. A particular care must be taken on the implementation for being efficient.

3.3. Experimental aspects

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 [26]. For inclined cylindrical holes, the experimental

database produced by Gustafsson and Johansson [2] represents a sound reference base and for normal injection, the work by [32] served as reference for LES simulations [30]. 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 [23]. In 2007, A. Most developed at UPPA a test facility for studying jet in crossflow issued from shaped holes [27]. 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 [28] and LES [29]. This test facility will be used in the framework of the present project to investigate a 1-hole plane i.e. an isolated jet in crossflow. PIV and LDV metrology will be used.

[2] http://www.tfd.chalmers.se/~gujo/WS11_2005/Slanted_jet/INDEX.HTM

CALVI Project-Team

3. Scientific Foundations

3.1. Kinetic models for plasma and beam physics

plasma physics, beam physics, kinetic models, reduced models, Vlasov equation, modeling, mathematical analysis, asymptotic analysis, existence, uniqueness

Plasmas and particle beams can be described by a hierarchy of models including N-body interaction, kinetic models and fluid models. Kinetic models in particular are posed in phase-space and involve specific difficulties. We perform a mathematical analysis of such models and try to find and justify approximate models using asymptotic analysis.

3.1.1. Models for plasma and beam physics

The **plasma state** can be considered as the **fourth state of matter**, obtained for example by bringing a gas to a very high temperature ($10^4\ K$ or more). The thermal energy of the molecules and atoms constituting the gas is then sufficient to start ionization when particles collide. A globally neutral gas of neutral and charged particles, called **plasma**, is then obtained. Intense charged particle beams, called nonneutral plasmas by some authors, obey similar physical laws.

The hierarchy of models describing the evolution of charged particles within a plasma or a particle beam includes N-body models where each particle interacts directly with all the others, kinetic models based on a statistical description of the particles and fluid models valid when the particles are at a thermodynamical equilibrium.

In a so-called *kinetic model*, each particle species s in a plasma or a particle beam is described by a distribution function $f_s(\mathbf{x}, \mathbf{v}, t)$ corresponding to the statistical average of the particle distribution in phase-space corresponding to many realisations of the physical system under investigation. The product $f_s\, d\mathbf{x}\, d\mathbf{v}$ is the average number of particles of the considered species, the position and velocity of which are located in a bin of volume $d\mathbf{x}\, d\mathbf{v}$ centered around $(\mathbf{x}, \mathbf{v})$. The distribution function contains a lot more information than what can be obtained from a fluid description, as it also includes information about the velocity distribution of the particles.

A kinetic description is necessary in collective plasmas where the distribution function is very different from the Maxwell-Boltzmann (or Maxwellian) distribution which corresponds to the thermodynamical equilibrium, otherwise a fluid description is generally sufficient. In the limit when collective effects are dominant with respect to binary collisions, the corresponding kinetic equation is the *Vlasov equation*

$$\frac{\partial f_s}{\partial t} + \mathbf{v} \cdot \frac{\partial f_s}{\partial \mathbf{x}} + \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \frac{\partial f_s}{\partial \mathbf{v}} = 0,$$

which expresses that the distribution function f is conserved along the particle trajectories which are determined by their motion in their mean electromagnetic field. The Vlasov equation which involves a self-consistent electromagnetic field needs to be coupled to the Maxwell equations in order to compute this field

$$\begin{aligned} -\frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} + \nabla \times \mathbf{B} &= \mu_0\, \mathbf{J}, \\ \frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} &= 0, \\ \div \mathbf{E} &= \frac{\rho}{\varepsilon_0}, \\ \div \mathbf{B} &= 0, \end{aligned}$$

which describes the evolution of the electromagnetic field generated by the charge density

$$\rho(\mathbf{x},t) = \sum_s q_s \int f_s(\mathbf{x},\mathbf{v},t)\, d\mathbf{v},$$

and current density

$$\mathbf{J}(\mathbf{x},t) = \sum_s q_s \int f_s(\mathbf{x},\mathbf{v},t)\mathbf{v}\, d\mathbf{v},$$

associated to the charged particles.

When binary particle-particle interactions are dominant with respect to the mean-field effects then the distribution function f obeys the Boltzmann equation

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\frac{\partial f}{\partial \mathbf{x}} = Q(f,f),$$

where Q is the nonlinear Boltzmann collision operator. In some intermediate cases, a collision operator needs to be added to the Vlasov equation.

The numerical solution of the three-dimensional Vlasov-Maxwell system represents a considerable challenge due to the huge size of the problem. Indeed, the Vlasov-Maxwell system is nonlinear and posed in phase space. It thus depends on seven variables: three configuration space variables, three velocity space variables and time, for each species of particles. This feature makes it essential to use every possible option to find a reduced model wherever possible, in particular when there are geometrical symmetries or small terms which can be neglected.

3.1.2. Mathematical and asymptotic analysis of kinetic models

The mathematical analysis of the Vlasov equation is essential for a thorough understanding of the model as well for physical as for numerical purposes. It has attracted many researchers since the end of the 1970s. Among the most important results which have been obtained, we can cite the existence of strong and weak solutions of the Vlasov-Poisson system by Horst and Hunze [74], see also Bardos and Degond [51]. The existence of a weak solution for the Vlasov-Maxwell system has been proved by Di Perna and Lions [59]. An overview of the theory is presented in a book by Glassey [71].

Many questions concerning for example uniqueness or existence of strong solutions for the three-dimensional Vlasov-Maxwell system are still open. Moreover, their is a realm of approached models that need to be investigated. In particular, the Vlasov-Darwin model for which we could recently prove the existence of global solutions for small initial data [52].

On the other hand, the asymptotic study of the Vlasov equation in different physical situations is important in order to find or justify reduced models. One situation of major importance in tokamaks, used for magnetic fusion as well as in atmospheric plasmas, is the case of a large external magnetic field used for confining the particles. The magnetic field tends to incurve the particle trajectories which eventually, when the magnetic field is large, are confined along the magnetic field lines. Moreover, when an electric field is present, the particles drift in a direction perpendicular to the magnetic and to the electric field. The new time scale linked to the cyclotron frequency, which is the frequency of rotation around the magnetic field lines, comes in addition to the other time scales present in the system like the plasma frequencies of the different particle species. Thus, many different time scales as well as length scales linked in particular to the different Debye length are present in the system. Depending on the effects that need to be studied, asymptotic techniques allow to find reduced models. In this spirit, in the case of large magnetic fields, recent results have been obtained by Golse and Saint-Raymond [72], [80] as well as by Brenier [57]. Our group has also contributed to this problem using homogenization techniques to justify the guiding center model and the finite Larmor radius model which are used by physicist in this setting [67], [65], [66].

Another important asymptotic problem yielding reduced models for the Vlasov-Maxwell system is the fluid limit of collisionless plasmas. In some specific physical situations, the infinite system of velocity moments of the Vlasov equations can be closed after a few of those, thus yielding fluid models.

3.2. Development of simulation tools

Numerical methods, Vlasov equation, unstructured grids, adaptivity, numerical analysis, convergence, Semi-Lagrangian method The development of efficient numerical methods is essential for the simulation of plasmas and beams. Indeed, kinetic models are posed in phase space and thus the number of dimensions is doubled. Our main effort lies in developing methods using a phase-space grid as opposed to particle methods. In order to make such methods efficient, it is essential to consider means for optimizing the number of mesh points. This is done through different adaptive strategies. In order to understand the methods, it is also important to perform their mathematical analysis. Since a few years we are interested also with solvers that uses Particle In Cell method. This new issue allows us to enrich some parts of our research activities previously centered on the Semi-Lagrangian approach.

3.2.1. Introduction

The numerical integration of the Vlasov equation is one of the key challenges of computational plasma physics. Since the early days of this discipline, an intensive work on this subject has produced many different numerical schemes. One of those, namely the Particle-In-Cell (PIC) technique, has been by far the most widely used. Indeed it belongs to the class of Monte Carlo particle methods which are independent of dimension and thus become very efficient when dimension increases which is the case of the Vlasov equation posed in phase space. However these methods converge slowly when the number of particles increases, hence if the complexity of grid based methods can be decreased, they can be the better choice in some situations. This is the reason why one of the main challenges we address is the development and analysis of adaptive grid methods.

3.2.2. Convergence analysis of numerical schemes

Exploring grid based methods for the Vlasov equation, it becomes obvious that they have different stability and accuracy properties. In order to fully understand what are the important features of a given scheme and how to derive schemes with the desired properties, it is essential to perform a thorough mathematical analysis of this scheme, investigating in particular its stability and convergence towards the exact solution.

3.2.3. The semi-Lagrangian method

The semi-Lagrangian method consists in computing a numerical approximation of the solution of the Vlasov equation on a phase space grid by using the property of the equation that the distribution function f is conserved along characteristics. More precisely, for any times s and t, we have

$$f(\mathbf{x},\mathbf{v},t) = f(\mathbf{X}(s;\mathbf{x},\mathbf{v},t),\mathbf{V}(s;\mathbf{x},\mathbf{v},t),s),$$

where $(\mathbf{X}(s;\mathbf{x},\mathbf{v},t),\mathbf{V}(s;\mathbf{x},\mathbf{v},t))$ are the characteristics of the Vlasov equation which are solution of the system of ordinary differential equations

$$\begin{aligned} \frac{d\mathbf{X}}{ds} &= \mathbf{V}, \\ \frac{d\mathbf{V}}{ds} &= \mathbf{E}(\mathbf{X}(s),s) + \mathbf{V}(s)\times\mathbf{B}(\mathbf{X}(s),s), \end{aligned} \tag{4}$$

with initial conditions $\mathbf{X}(t)=\mathbf{x}$, $\mathbf{V}(t)=\mathbf{v}$.

From this property, f^n being known one can induce a numerical method for computing the distribution function f^{n+1} at the grid points $(\mathbf{x}_i, \mathbf{v}_j)$ consisting in the following two steps:

1. For all i, j, compute the origin of the characteristic ending at $\mathbf{x}_i, \mathbf{v}_j$, i.e. an approximation of $\mathbf{X}(t_n; \mathbf{x}_i, \mathbf{v}_j, t_{n+1})$, $\mathbf{V}(t_n; \mathbf{x}_i, \mathbf{v}_j, t_{n+1})$.
2. As

$$f^{n+1}(\mathbf{x}_i, \mathbf{v}_j) = f^n(\mathbf{X}(t_n; \mathbf{x}_i, \mathbf{v}_j, t_{n+1}), \mathbf{V}(t_n; \mathbf{x}_i, \mathbf{v}_j, t_{n+1})),$$

   f^{n+1} can be computed by interpolating f^n which is known at the grid points at the points $\mathbf{X}(t_n; \mathbf{x}_i, \mathbf{v}_j, t_{n+1}), \mathbf{V}(t_n; \mathbf{x}_i, \mathbf{v}_j, t_{n+1})$.

This method can be simplified by performing a time-splitting separating the advection phases in physical space and velocity space, as in this case the characteristics can be solved explicitly.

3.2.4. Adaptive semi-Lagrangian methods

Uniform meshes are most of the time not efficient to solve a problem in plasma physics or beam physics as the distribution of particles is evolving a lot as well in space as in time during the simulation. In order to get optimal complexity, it is essential to use meshes that are fitted to the actual distribution of particles. If the global distribution is not uniform in space but remains locally mostly the same in time, one possible approach could be to use an unstructured mesh of phase space which allows to put the grid points as desired. Another idea, if the distribution evolves a lot in time is to use a different grid at each time step which is easily feasible with a semi-Lagrangian method. And finally, the most complex and powerful method is to use a fully adaptive mesh which evolves locally according to variations of the distribution function in time. The evolution can be based on a posteriori estimates or on multi-resolution techniques.

3.2.5. Particle-In-Cell codes

The Particle-In-Cell method [56] consists in solving the Vlasov equation using a particle method, i.e. advancing numerically the particle trajectories which are the characteristics of the Vlasov equation, using the equations of motion which are the ordinary differential equations defining the characteristics. The self-fields are computed using a standard method on a structured or unstructured grid of physical space. The coupling between the field solve and the particle advance is done on the one hand by depositing the particle data on the grid to get the charge and current densities for Maxwell's equations and, on the other hand, by interpolating the fields at the particle positions. This coupling is one of the difficult issues and needs to be handled carefully.

3.2.6. Maxwell's equations in singular geometry

The solutions to Maxwell's equations are *a priori* defined in a function space such that the curl and the divergence are square integrable and that satisfy the electric and magnetic boundary conditions. Those solutions are in fact smoother (all the derivatives are square integrable) when the boundary of the domain is smooth or convex. This is no longer true when the domain exhibits non-convex *geometrical singularities* (corners, vertices or edges).

Physically, the electromagnetic field tends to infinity in the neighbourhood of the re-entrant singularities, which is a challenge to the usual finite element methods. Nodal elements cannot converge towards the physical solution. Edge elements demand considerable mesh refinement in order to represent those infinities, which is not only time- and memory-consuming, but potentially catastrophic when solving time dependent equations: the CFL condition then imposes a very small time step. Moreover, the fields computed by edge elements are discontinuous, which can create considerable numerical noise when the Maxwell solver is embedded in a plasma (e.g. PIC) code.

In order to overcome this dilemma, a method consists in splitting the solution as the sum of a *regular* part, computed by nodal elements, and a *singular* part which we relate to singular solutions of the Laplace operator, thus allowing to calculate a local analytic representation. This makes it possible to compute the solution precisely without having to refine the mesh.

This *Singular Complement Method* (SCM) had been developed [49] and implemented [48] in plane geometry.

An especially interesting case is axisymmetric geometry. This is still a 2D geometry, but more realistic than the plane case; despite its practical interest, it had been subject to much fewer theoretical studies [54]. The non-density result for regular fields was proven [58], the singularities of the electromagnetic field were related to that of modified Laplacians [45], and expressions of the singular fields were calculated [46]. Thus the SCM was extended to this geometry. It was then implemented by F. Assous (now at Bar-Ilan University, Israel) and S. Labrunie in a PIC–finite element Vlasov–Maxwell code [47].

As a byproduct, space-time regularity results were obtained for the solution to time-dependent Maxwell's equation in presence of geometrical singularities in the plane and axisymmetric cases [70], [46].

3.3. Large size problems

Parallelism, domain decomposition, code transformation

3.3.1. Introduction

The applications we consider lead to very large size computational problems for which we need to apply modern computing techniques enabling to use efficiently many computers including traditional high performance parallel computers and computational grids.

The full Vlasov-Maxwell system yields a very large computational problem mostly because the Vlasov equation is posed in six-dimensional phase-space. In order to tackle the most realistic possible physical problems, it is important to use all the modern computing power and techniques, in particular parallelism and grid computing.

3.3.2. Parallelization of numerical methods

An important issue for the practical use of the methods we develop is their parallelization. We address the problem of tuning these methods to homogeneous or heterogeneous architectures with the aim of meeting increasing computing resources requirements.

Most of the considered numerical methods apply a series of operations identically to all elements of a geometric data structure: the mesh of phase space. Therefore these methods intrinsically can be viewed as a data-parallel algorithm. A major advantage of this data-parallel approach derives from its scalability. Because operations may be applied identically to many data items in parallel, the amount of parallelism is dictated by the problem size.

Parallelism, for such data-parallel PDE solvers, is achieved by partitioning the mesh and mapping the submeshes onto the processors of a parallel architecture. A good partition balances the workload while minimizing the communications overhead. Many interesting heuristics have been proposed to compute near-optimal partitions of a (regular or irregular) mesh. For instance, the heuristics based on space-filing curves [73] give very good results for a very low cost.

Adaptive methods include a mesh refinement step and can highly reduce memory usage and computation volume. As a result, they induce a load imbalance and require to dynamically distribute the adaptive mesh. A problem is then to combine distribution and resolution components of the adaptive methods with the aim of minimizing communications. Data locality expression is of major importance for solving such problems. We use our experience of data-parallelism and the underlying concepts for expressing data locality [81], optimizing the considered methods and specifying new data-parallel algorithms.

As a general rule, the complexity of adaptive methods requires to define software abstractions allowing to separate/integrate the various components of the considered numerical methods (see [79] as an example of such modular software infrastructure).

Another key point is the joint use of heterogeneous architectures and adaptive meshes. It requires to develop new algorithms which include new load balancing techniques. In that case, it may be interesting to combine several parallel programming paradigms, i.e. data-parallelism with other lower-level ones.

Moreover, exploiting heterogeneous architectures requires the use of a run time support associated with a programming interface that enables some low-level hardware characteristics to be unified. Such run time support is the basis for heterogeneous algorithmics. Candidates for such a run time support may be specific implementations of MPI such as MPICH-G2 (a grid-enabled MPI implementation on top of the GLOBUS tool kit for grid computing [64]).

Our general approach for designing efficient parallel algorithms is to define code transformations at any level. These transformations can be used to incrementally tune codes to a target architecture and they warrant code reusability.

CASTOR Team

3. Scientific Foundations

3.1. Plasma Physics

Participants: Hervé Guillard, Boniface Nkonga, Afeintou Sangam, Richard Pasquetti, Audrey Bonnement, Marie Martin, Cédric Lachat, Laure Combe, Jacques Blum, Cédric Boulbe, Sebastian Minjeaud.

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: Alain Dervieux, 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: 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 [1] 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.

[1] 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

CLASSIC Project-Team

3. Scientific Foundations

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]. It has introduced the idea of combining the PAC-Bayesian approach with the use of thresholded exponential moments, in order to derive bounds under very weak assumptions on the noise.

3.1.2. Sparsity and ℓ_1–regularization

Another line of research in regression estimation is the use of sparse models, and its link with ℓ_1–regularization. Regularization is the joint minimization of some empirical criterion and some penalty function; it should lead to a model that not only fits well the data but is also as simple as possible.

For instance, the Lasso uses a ℓ^1–regularization instead of a ℓ^0–one; it is popular mostly because it leads to *sparse* solutions (the estimate has only a few nonzero coordinates), which usually have a clear interpretation in many settings (e.g., the influence or lack of influence of some variables). In addition, unlike ℓ^0–penalization, the Lasso is *computationally feasible* for high-dimensional data.

3.1.3. Pushing it to the extreme: no assumption on the data

The next brick of our scientific foundations explains why and how, in certains cases, we may formulate absolutely no assumption on the data (x_i, y_i), $i = 1, \cdots, n$, which is then considered a deterministic set of input–output pairs.

3.2. On-line aggregation of predictors for the prediction of time series, with or without stationarity assumptions

We are concerned here with *sequential prediction* of outcomes, given some base predictions formed by *experts*. We distinguish two settings, depending on how the sequence of outcomes is generated: it is either

- the realization of some stationary process,
- or is not modeled at all as the realization of any underlying stochastic process (these sequences are called *individual sequences*).

The aim is to predict almost as well as the best expert. Typical good forecasters maintain one weight per expert, update these weights depending on the past performances, and output at each step the corresponding weighted linear combination of experts' advices.

The difference between the cumulative prediction error of the forecaster and the one of the best expert is called the regret. The goal here is to upper bound the regret by a quantity as small as possible.

3.3. Multi-armed bandit problems, prediction with limited feedback

We are interested in settings in which the feedback obtained on the predictions is limited, in the sense that it does not fully reveal what actually happened.

3.3.1. Bandit problems

This is also a sequential problem in which some regret is to be minimized.

However, this problem is a stochastic problem: a large number of arms, possibly indexed by a continuous set like $[0, 1]$, is available. Each arm is associated with a fixed but unknown distribution. At each round, the player chooses an arm, a payoff is drawn at random according to the distribution that is associated with it, and the only feedback that the player gets is the value of this payoff. The key quantity to study this problem is the mean-payoff function f, that indicates for each arm x the expected payoff $f(x)$ of the distribution that is associated with it. The target is to minimize the regret, i.e., ensure that the difference between the cumulative payoff obtained by the player and the one of the best arm is small.

3.3.2. A generalization of the regret: the approachability of sets

Approachability is the ability to control random walks. At each round, a vector payoff is obtained by the first player, depending on his action and on the action of the opponent player. The aim is to ensure that the average of the vector payoffs converges to some convex set. Necessary and sufficient conditions were obtained by Blackwell and others to ensure that such strategies exist, both in the full information and in the bandit cases.

Some of these results can be extended to the case of games with signals (games with partial monitoring), where at each round the only feedback obtained by the first player is a random signal drawn according to a distribution that depends on the action profile taken by the two players, while the opponent player still has a full monitoring.

COFFEE Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

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.

COMMANDS Project-Team

3. Scientific Foundations

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 [69]), with improvements due to the "Chicago school", Bliss [44] during the first part of the 20th century, and by the notion of relaxed problem and generalized solution (Young [77]).

Trajectory optimization really started with the spectacular achievement done by Pontryagin's group [75] 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 [50], Leitmann [71], Lee and Markus [70], Ioffe and Tihomirov [66]). Since then, various theoretical achievements have been obtained by extending the results to nonsmooth problems, see Aubin [40], Clarke [51], Ekeland [58].

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 [53], [54], [52]. 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 [42] and Bardi and Capuzzo-Dolcetta [41].

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 [62], Bonnans [47]. 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. [57]. See also Malanowski et al. [72].

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 [45], [46], [39]. As initiated in [61], 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 [53], [54], [52], 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 [60], [59] 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 [63], [64], [65], [73], and also the schemes on nonregular grids by R. Abgrall [38], [37]. 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 [76] and recently revisited by Després-Lagoutière [56], [55]. 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 [78]) 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 [74], that happens also to be well adapted to the study of discretization errors for numerical discretization schemes [67], [43].

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 [68] 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 [49]. 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 [48].

CONCHA Project-Team

3. Scientific Foundations

3.1. Challenges related to numerical simulations of complex flows

First, we describe some typical difficulties in our fields of application which require the improvement of established and the development of new methods.

- Coupling of equations and models
  The general equations of fluid dynamics consist in a strongly coupled nonlinear system. Its mathematical nature depends on the precise model, but in general contains hyperbolic, parabolic, and elliptic parts. The spectrum of physical phenomena described by these equations is very large: convection, diffusion, waves... In addition, it is often necessary to couple different models in order to describe different parts of a mechanical system: chemistry, fluid-fluid-interaction, fluid-solid-interaction...
- Robustness with respect to physical parameters
  The values of physical parameters such as diffusion coefficients and constants describing different state equations and material laws lead to different behaviour characterized for example by the Reynolds, Mach, and Weissenberg numbers. Optimized numerical methods are available in many situations, but it remains a challenging problem in some fields of applications to develop robust discretizations and solution algorithms.
- Multiscale phenomena
  The inherent nonlinearities lead to an interplay of a wide range of physical modes, well-known for example from the study of turbulent flows. Since the resolution of all modes is often unreachable, it is a challenging task to develop numerical methods, which are still able to reproduce the essential features of the physical phenomenon under study.

3.2. Stabilized and discontinuous finite element methods

The discontinuous Galerkin method [68], [66], [44], [43] has gained enormous success in CFD due to its flexibility, links with finite volume methods, and its local conservation properties. In particular, it seems to be the most widely used finite element method for the Euler equations [45]. On the other hand, the main drawback of this approach is the large number of unknowns as compared to standard finite element methods. The situation is even worse if one counts the population of the resulting system matrices. In order to find a more efficient approach, it seems therefore important to study the connections with other finite element methods.

In view of the ubiquous problem of large Péclet numbers, stabilization techniques have been introduced since a long time. They are either based on upwinding or additional terms in the discrete variational formulation. The drawback of the first technique is a loss in consistency which generally leads to large numerical diffusion. The grand-father of the second technique is the SUPG/GLS method [57], [67]. Recently, new approaches have been developed, which try do avoid coupling of the different equations due to the residuals. In this context we cite LPS (local projection stabilization) [62], [55], [48][5] and CIP (continuous interior penalty) [58], [59].

3.3. Finite element methods on quadrilateral and hexahedral meshes

The construction of finite element methods on quadrilateral, and particularly, hexahedral meshes can be a complicated task; especially the development of mixed and non-conforming methods is an active field of research. The difficulties arise not only from the fact that adequate degrees of freedom have to be found, but also from the non-constantness of the element Jacobians; an arbitrary hexahedron, which we define as the image of the unit cube under a tri-linear transformation, does in general not have plane faces, which implies for example, that the normal vector is not constant on a side.

In collaboration with Eric Dubach (Associate professor at LMAP) and Jean-Marie Thomas (Former professor at LMAP) we have built a new class of finite element functions (named pseudo-conforming) on quadrilateral and hexahedral meshes. The degrees of freedom are the same as those of classical iso-parametric finite elements but the basis functions are defined as polynomials on each element of the mesh. On general quadrilaterals and hexahedra, our method leads to a non-conforming method; in the particular case of parallelotopes, the new finite elements coincide with the classical ones [61], [60].

3.4. Finite element methods for interface problems

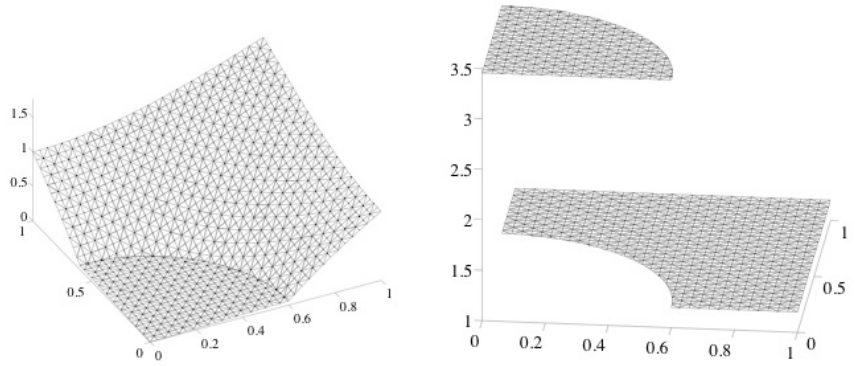


Figure 1. Incompressible elasticity with discontinuous material properties (left: modulus of velocities, right: pressure; from [46]).

The NXFEM (Nitsche eXtended finite element method) has been developed in [63] and [64]. It is based on a pure variational formulation with standard finite element spaces, which are locally enriched in such a way that the accurate capturing of an interface not aligned with the underlying mesh is possible, giving a rigorous formulation of the very popular XFEM. A typical computation for the Stokes problem with varying, piecewise constant viscosity is shown in Figure 1 . This technology opens the door to many applications in the field of fluid mechanics, such as immiscible flows, free surface flows and so on.

3.5. Adaptivity

Adaptive finite element methods are becoming a standard tool in numerical simulations, and their application in CFD is one of the main topics of Concha. Such methods are based on a posteriori error estimates of the discretization error avoiding explicit knowledge of properties of the solution, in contrast to a priori error estimates. The estimator is used in an adaptive loop by means of a local mesh refinement algorithm. The mathematical theory of these algorithms has for a long time been bounded to the proof of upper and lower bounds, but has made important improvements in recent years. For illustration, a typical sequence of adaptively refined meshes on an L-shaped domain is shown in Figure 2 .

The theoretical analysis of mesh-adaptive methods, even in the most standard case of the Poisson problem, is in its infancy. The first important results in this direction concern the convergence of the sequence of solution generated by the algorithm (the standard a priori error analysis does not apply since the global mesh-size does not necessarily go to zero). In order to prove convergence, an unavoidable data approximation term has to be treated in addition to the error estimator [69]. These result do not say anything about the convergence speed, that is the number of unknowns required to achieve a given accuracy. Such complexity estimates are the subject of active research, the first fundamental result in this direction is [54].

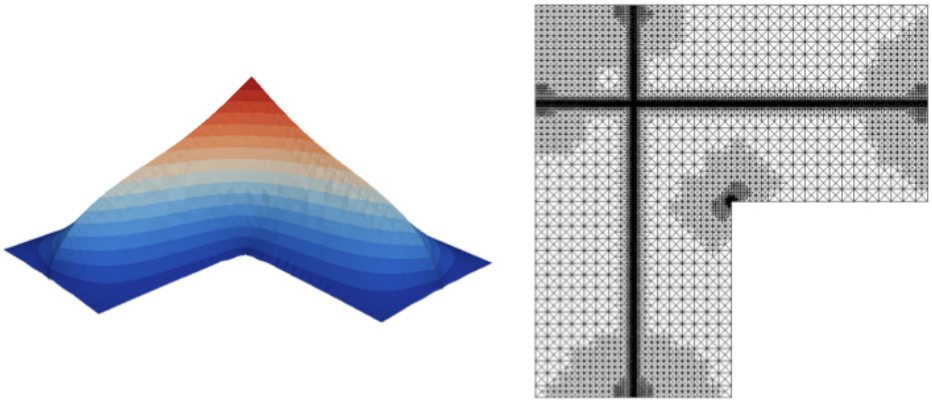


Figure 2. Solution with rough right-hand-side in a corner domain and adaptively refined mesh (from [50]).

Our first contribution [23] to this field has been the introduction of a new adaptive algorithm which makes use of an adaptive marking strategy, which refines according to the data oscillations only if they are by a certain factor larger then the estimator. This algorithm allowed us to prove geometric convergence and quasi-optimal complexity, avoiding additional iteration as used before [71]. We have extended our results to conforming FE without inner node refinement [51] and to mixed FE [50]. In this case, a major additional difficulty arises from the fact that, due to the saddle-point formulation, the orthogonality relation known from continuous FEM does not hold. In addition, we have considered the case of incomplete solution of the discrete systems. To this end, we have developed a simple adaptive stopping criterion based on comparison of the iteration error with the discretization error estimator, see also [49].

Goal-oriented error estimation has been introduced in [52]. It allows to error control and adaptivity directly oriented to the computation of physical quantities, such as the drag and lift coefficient, the Nusselt number, and other physical quantities.

CORIDA Project-Team

3. Scientific Foundations

3.1. Analysis and control of fluids and of fluid-structure interactions

Participants: Thomas Chambrion, Antoine Henrot, Alexandre Munnier, Lionel Rosier, Jean-François Scheid, Takeo 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 [59] and the references therein). The feedback control of fluid motion has also been recently investigated by several research teams (see, for instance Barbu [54] 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 (see [8]). Another important result gives the existence and the uniqueness of strong solutions for two or three-dimensional Navier–Stokes fluid (see [9]). 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, Bertrand Thierry.

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 [63]. This characterization can be related to the Hautus criterion in the theory of finite dimensional systems (cf. [58]). 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 [64]. Using the result of Miller, K. Ramdani, T. Takahashi, M. Tucsnak have obtained in [5] a new spectral formulation of the criterion of Liu [63], 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 [56] 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 [57]. 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, 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 [61] , 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 [66]. In the case of systems governed by linear PDE's some similar estimates have been obtained only very recently (see, for instance Miller [64]). 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 [53]).

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 [67], for optimal and robust control problems.

CQFD Project-Team

3. Scientific Foundations

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 [88] and Duan and Li [76]
  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 [77]
- Stochastic optimal control: optimal stopping, impulse control, continuous control, linear programming, singular perturbation, martingale problem.
  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.

Analysis of various problems arising in MDPs leads to a large variety of interesting mathematical problems. The second objective of the team is to study some theoretical aspects related to MDPs such as convex analytical methods and singular perturbation.

DEFI Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

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 (tological) 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, echography, 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 muli-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 caretisian 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 clalled "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 of 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) numercially 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. Scientific Foundations

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], [85], [106], [87], [90].

- 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. [98]). 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], [31].

- 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. Scientific Foundations

3.1. Modeling and landscape analysis

The modeling of problems, the analysis of structures (landscapes) of MOPs and the performance assessment of resolution methods are significant topics in the design of optimization methods. The effectiveness of metaheuristics depends on the properties of the problem and its landscape (roughness, convexity, etc). The notion of landscape has been first described in [89] by the way of the study of species evolution. Then, this notion has been used to analyze combinatorial optimization problems.

3.1.1. Modeling of problems

Generally there are several ways of modeling a given problem. First, one has to find the most suitable model for the type of resolution he or she plans to use. The choice can be made after a theoretical analysis of the model, or after computational experiments. The choice of the model depends on the type of method used. For example, a major issue in the design of exact methods is to find tight relaxations for the problem considered.

Let us note that many combinatorial optimization problems of the literature have been studied in their mono-objective form even if a lot of them are naturally of a multi-objective nature.

Therefore, in the DOLPHIN project, we address the modeling of MOPs in two phases. The first one consists in studying the mono-objective version of the problem, where all objectives but one are considered as constraints. In the second phase, we propose methods to adapt the mono-objective models or to create hand-tailored models for the multi-objective case. The models used may come from the first phase, or from the literature.

3.1.2. Analysis of the structure of a problem

The landscape is defined by a neighborhood operator and can be represented by a graph $G = (V, E)$. The vertices represent the solutions of the problem and an edge (e_1, e_2) exists if the solution e_2 can be obtained by an application of the neighborhood operator on the solution e_1. Then, considering this graph as the ground floor, we elevate each solution to an altitude equals to its cost. We obtain a surface, or landscape, made of peaks, valleys, plateaus, cliffs, etc. The problem lies in the difficulty to have a realistic view of this landscape.

Like others, we believe that the main point of interest in the domain of combinatorial optimization is not the design of the best algorithm for a large number of problems but the search for the most adapted method to an instance or a set of instances of a given problem. Therefore, we are convinced that no ideal metaheuristic, designed as a black-box, may exist.

Indeed, the first studies realized in our research group on the analysis of landscapes of different mono-objective combinatorial optimization problems (traveling salesman problem, quadratic assignment problem) have shown that not only different problems correspond to different structures but also that different instances of the same problem correspond to different structures.

For instance, we have realized a statistical study of the landscapes of the quadratic assignment problem. Some indicators that characterize the landscape of an instance have been proposed and a taxonomy of the instances including three classes has been deduced. Hence it is not enough to adapt the method to the problem under study but it is necessary to specialize it according to the type of the treated instance.

So in its studies of mono-objective problems, the DOLPHIN research group has introduced into the resolution methods some information about the problem to be solved. The landscapes of some combinatorial problems have been studied in order to investigate the intrinsic natures of their instances. The resulting information has been inserted into an optimization strategy and has allowed the design of efficient and robust hybrid methods. The extension of these studies to multi-objective problems is a part of the DOLPHIN project [87], [88].

3.1.3. Performance assessment

The DOLPHIN project is also interested in the performance assessment of multi-objective optimization methods. Nowadays, statistical techniques developed for mono-objective problems can be adapted to the multi-objective case. Nevertheless, specific tools are necessary in many situations: for example, the comparison of two different algorithms is relatively easy in the mono-objective case - we compare the quality of the best solution obtained in a fixed time, or the time needed to obtain a solution of a certain quality. The same idea cannot be immediately transposed to the case where the output of the algorithms is a set of solutions having several quality measures, and not a single solution.

Various indicators have been proposed in the literature for evaluating the performance of multi-objective optimization methods but no indicator seems to outperform the others [90]. The DOLPHIN research group has proposed two indicators: the *contribution* and the *entropy* [82]. The contribution evaluates the supply in term of Pareto-optimal solutions of a front compared to another one. The entropy gives an idea of the diversity of the solutions found. These two metrics are used to compare the different metaheuristics in the research group, for example in the resolution of the bi-objective flow-shop problem, and also to show the contribution of the various mechanisms introduced in these metaheuristics.

3.1.4. Goals

One of the main issues in the DOLPHIN project is the study of the landscape of multi-objective problems and the performance assessment of multi-objective optimization methods to design efficient and robust resolution methods:

- *Landscape study:* The goal here is to extend the study of landscapes of the mono-objective combinatorial optimization problems to multi-objective problems in order to determine the structure of the Pareto frontier and to integrate this knowledge about the problem structure in the design of resolution methods.

  This study has been initiated for the bi-objective flow-shop problem. We have studied the convexity of the frontiers obtained in order to show the interest of our Pareto approach compared to an aggregation approach, which only allows one to obtain the Pareto solutions situated on the convex hull of the Pareto front (supported solutions).

  Our preliminary study of the landscape of the bi-objective flow-shop problem shows that the supported solutions are very closed to each other. This remark leads us to improve an exact method initially proposed for bi-objective problems. Furthermore, a new exact method able to deal with any number of objectives has been designed.
- *Performance assessment: The goal here is to extend GUIMOO in order to provide efficient visual and metric tools for evaluating the assessment of multi-objective resolution methods.*

3.2. Hybrid multi-objective optimization methods

The success of metaheuristics is based on their ability to find efficient solutions in a reasonable time [76]. 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 [59], [74]. To be convinced one may refer to the list of references on Evolutionary Multi-objective Optimization maintained by Carlos A. Coello Coello [4], which contains more

[4] http://www.lania.mx/~ccoello/EMOO/EMOObib.html

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'.

3.2.1. 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 [84]. 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.
- *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 if 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.

3.2.2. 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 [68].

3.2.3. 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).

3.3. 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 goal 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.

3.3.1. 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 [73], [83].

3.3.2. 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 architecture 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.

GAMMA3 Project-Team (section vide)

GECO Team

3. Scientific Foundations

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... [27], [62] 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 ([83], [70], [76]), those exploiting the possible flatness of the system ([56]) and those based on the continuation method ([95]). 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 ([82]), geometric measure theory ([57], [31]) and hypoelliptic operator theory ([43]).

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 [49]).

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 [28] about the controllability of Navier–Stokes equation by low forcing modes.

GEOSTAT Project-Team

3. Scientific Foundations

3.1. Dynamics of complex systems

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. Linear methods do not unlock the multiscale structures and cascading variables of primarily importance as previewed 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 [38] etc. have appeared during the last decades. 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 gazeous 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 [29], [26], [43], [34]. Some quantities related to nonlinearity, such as Lyapunov exponents, Kolmogorov-Sinai entropy etc. can be computed at least in the phase space [27]. 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 [39] about strange attractors in turbulence has motivated the determination of discrete dynamical systems associated to time series [31], 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 [28] and multiscale/multiresolution are the subjects of intense research and are profoundly reshaping the analysis of complex signals by nonlinear approaches [25], [30]. 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) [8] [5]. 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 [40] is local and reconstruction in the signal's domain is related to predictability.

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 $\log R_\tau(t)$:

$$\lambda_1 = \lim_{t\to\infty} \frac{1}{t} \langle \log R_\tau(t) \rangle \tag{5}$$

with $\langle\cdot\rangle$ being time average and R_τ the response to a perturbation [27]. More refined information is provided by the Kolmogorov-Sinai entropy:

$$h_{KS} = \lim_{\varepsilon\to 0} \lim_{t\to\infty} \frac{1}{t} \log N_{\text{eff}}(\varepsilon, t) \tag{6}$$

($N_{\text{eff}}(\varepsilon, t)$ is related to events which appear with very high probability in long time). In GEOSTAT our aim is to relate these classical quantities (among others) to the behaviour of SEs, which are defined by a limiting behaviour

$$\mu\left(\mathcal{B}_r(\mathbf{x})\right) \;=\; \alpha(\mathbf{x})\, r^{d+h(\mathbf{x})} \;+\; o\left(r^{d+h(\mathbf{x})}\right) \;\; (r \to 0) \tag{7}$$

(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:

$$\mathcal{T}_\Psi \mu(\mathbf{x}, r) = \int_{\mathbb{R}^d} \mathrm{d}\mu(\mathbf{x}') \frac{1}{r^d} \Psi\left(\frac{\mathbf{x} - \mathbf{x}'}{r}\right) \tag{8}$$

(Ψ: mother wavelet, admissible or not). SEs are related to the framework of reconstructible systems, and consequently to predictability. They unlock the geometric localization of multiscale structures in a complex signal:

$$\mathcal{F}_h = \{\mathbf{x} \in \Omega \;\mid\; h(\mathbf{x}) = h\}, \tag{9}$$

(Ω: signal's domain) and are consequently in relation with *optimal wavelets*:

$$\mathcal{T}_\psi[\mathbf{s}](\mathbf{x}, \mathbf{r_1}) = \zeta_{r_1/r_2}(\mathbf{x}) \mathcal{T}_\psi[\mathbf{s}](\mathbf{x}, \mathbf{r_2}) \tag{10}$$

($\mathbf{r_1} < \mathbf{r_2}$: two scales of observation, ζ: injection variable between the scales, ψ: optimal wavelet) and their **multiresolution analysis**. They are related to persistence along the scales and lead to multiresolution analysis whose coefficients verify

$$\alpha_s = \eta_1 \alpha_f + \eta_2 \tag{11}$$

with α_s and α_f referrin 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.

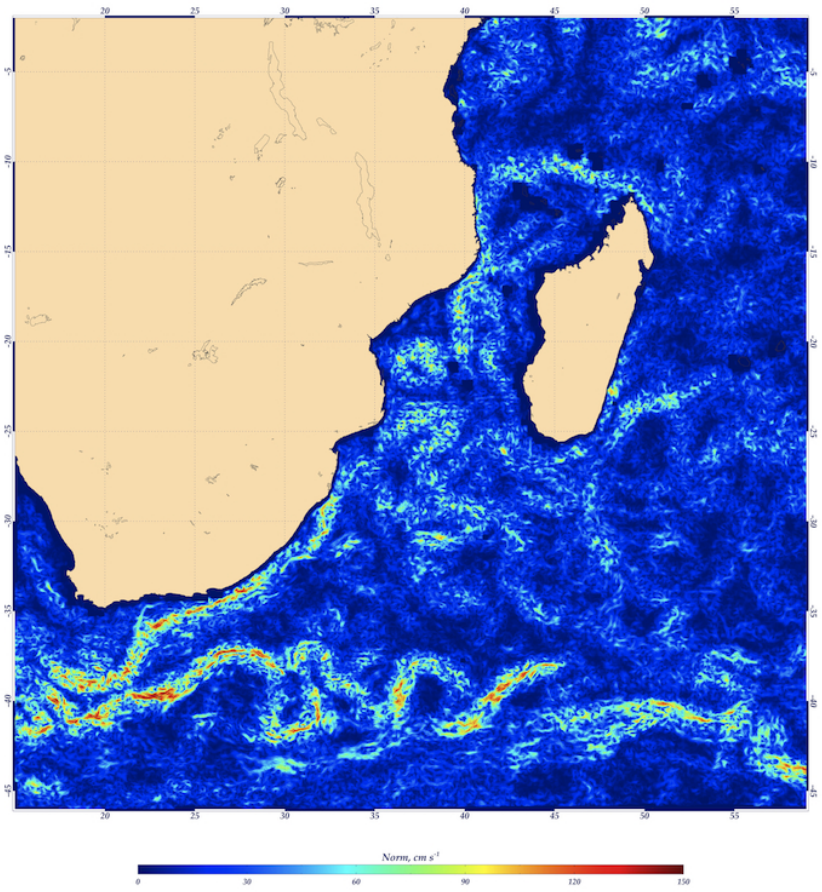


Figure 1. Visualization of the motion field computed at high spatial resolution (pixel size: 4kms) over a wide area around South Africa. The ocean dynamics is computed by propagating low resolution information coming from altimetry data (pixel size: 24 kms) along approximated optimal multiresolution analysis computed over the singularity exponents of Sea Surface Temperature data obtained from MODIS AQUA and OSTIA. Common work between GEOSTAT and DYNBIO (LEGOS, CNRS UMR 55 66, Toulouse).

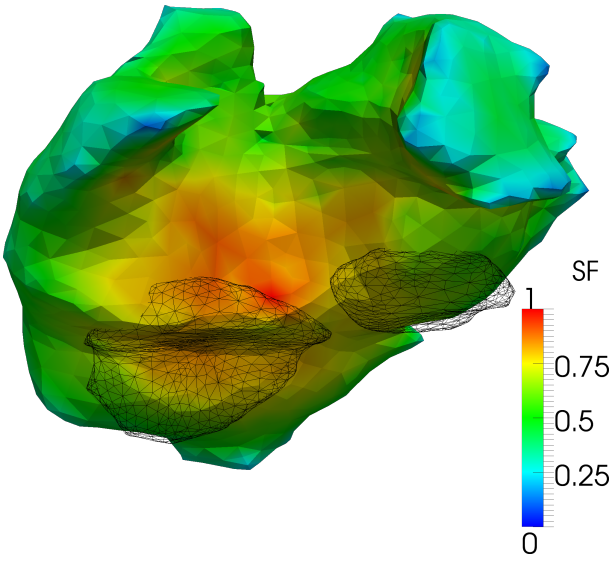


Figure 2. Result of the computation of a normalized source field over the 3D epicardial surface of the atria, from electric potential data acquired on a regular grid of electrodes placed on a patient's chest. There is a strong correlation between the red parts of the source field and the locations inside the heart where fibrillation occurs. Inputted data courtesy of IHU LIRYC.

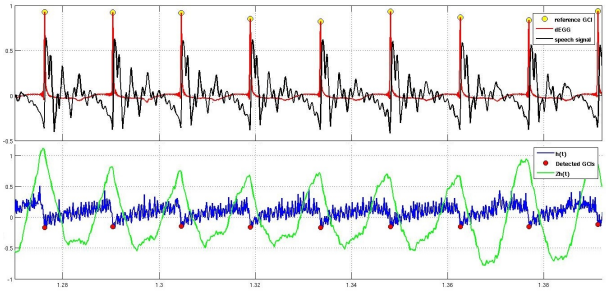


Figure 3. Top: A segment of a voiced speech signal (in black) along with the differentiated EGG (dEGG) recording (in red). Local maxima of dEGG shows the reference GCIs (yellow circles). Bottom: The singularity exponents (in blue) along with an auxiliary functional (in green) defined as $Zh(t) = \sum_{u=t-T_L}^{t-\delta t} h(u) - \sum_{u=t}^{t+T_L} h(u)$ *($h(t)$: singularity exponent at t). In each positive half-period of $Zh(t)$, the minimum of singularity exponents is taken as the GCI (red circle).*

In a first example we give some insight about the collaboration with LEGOS Dynbio team [1] 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 is introduced; 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. See figure 1 .

In a second example, we show in figure 2 the highly promising results 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 arythmic phenomena inside the heart [6].

Our last example is about speech. 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. See figure 3 .

[1] http://www.legos.obs-mip.fr/recherches/equipes/dynbio.

I4S Team

3. Scientific Foundations

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 [34] :

$$\widehat{\theta}_N=\arg\{\theta\in\Theta\,:\,\mathcal{K}_N(\theta)=0\}$$

Assuming that θ^* is the true parameter value, and that $\mathbf{E}_{\theta^*}[\mathcal{K}_N(\theta)]=0$ if and only if $\theta=\theta^*$ with θ^* fixed (identifiability condition), then $\widehat{\theta}_N$ converges towards θ^*. Thanks to the central limit theorem, the vector $\mathcal{K}_N(\theta^*)$ is asymptotically Gaussian with zero mean, with covariance matrix Σ which can be either computed or estimated. If, additionally, the matrix $\mathcal{J}_N=-\mathbf{E}_{\theta^*}[\mathcal{K}'_N(\theta^*)]$ is invertible, then using a Taylor expansion and the constraint $\mathcal{K}_N(\widehat{\theta}_N)=0$, the asymptotic normality of the estimate is obtained :

$$\sqrt{N}\,(\widehat{\theta}_N-\theta^*)\approx\mathcal{J}_N^{-1}\,\sqrt{N}\,\mathcal{K}_N(\theta^*)$$

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 explained in module 4.4 .

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 :

$$\text{(Safe) } \mathbf{H}_0 :\ \theta=\theta_0 \quad \text{ and } \quad \text{(Damaged) } \mathbf{H}_1 :\ \theta=\theta_0+\eta/\sqrt{N} \tag{12}$$

where η is an unknown but fixed change vector. A residual is generated under the form :

$$\zeta_N = 1/\sqrt{N}\ \sum_{k=0}^{N} K(\theta_0, Z_k) = \sqrt{N}\ \mathcal{K}_N(\theta_0)\ . \tag{13}$$

If the matrix $\mathcal{J}_N = -\mathbf{E}_{\theta_0}[\mathcal{K}'_N(\theta_0)]$ converges towards a limit $\mathcal{J}$, then the central limit theorem shows [31] that the residual is asymptotically Gaussian :

$$\zeta_N \xrightarrow[N \to \infty]{} \begin{cases} \mathcal{N}(0, \Sigma) & \text{under } \mathbf{P}_{\theta_0} \ , \\ \\ \mathcal{N}(\mathcal{J}\,\eta, \Sigma) & \text{under } \mathbf{P}_{\theta_0 + \eta/\sqrt{N}} \ , \end{cases} \tag{14}$$

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 :

$$\chi^2 = \overline{\zeta}^{\,T}\, \mathbf{F}^{-1}\, \overline{\zeta} \gtrless \lambda \ . \tag{15}$$

where

$$\overline{\zeta} \triangleq \mathcal{J}^T\, \Sigma^{-1}\, \zeta_N \quad \text{and} \quad \mathbf{F} \triangleq \mathcal{J}^T\, \Sigma^{-1}\, \mathcal{J} \tag{16}$$

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.

3.4.1. Isolation.

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 [29]. Here we only sketch two intuitively simple statistical nuisance elimination techniques, which proceed by projection and rejection, respectively.

The fault vector η is partitioned into an informative part and a nuisance part, and the sensitivity matrix $\mathcal{J}$, the Fisher information matrix $\mathbf{F} = \mathcal{J}^T\, \Sigma^{-1}\, \mathcal{J}$ and the normalized residual $\overline{\zeta} = \mathcal{J}^T\, \Sigma^{-1}\, \zeta_N$ are partitioned accordingly

$$\eta = \begin{pmatrix} \eta_a \\ \eta_b \end{pmatrix} \ , \quad \mathcal{J} = \begin{pmatrix} \mathcal{J}_a & \mathcal{J}_b \end{pmatrix} \ , \quad \mathbf{F} = \begin{pmatrix} \mathbf{F}_{aa} & \mathbf{F}_{ab} \\ \mathbf{F}_{ba} & \mathbf{F}_{bb} \end{pmatrix} \ , \quad \overline{\zeta} = \begin{pmatrix} \overline{\zeta}_a \\ \overline{\zeta}_b \end{pmatrix} \ .$$

A rather intuitive statistical solution to the isolation problem, which can be called *sensitivity* approach, consists in projecting the deviations in η onto the subspace generated by the components η_a to be isolated, and deciding between $\eta_a = \eta_b = 0$ and $\eta_a \neq 0$, $\eta_b = 0$. This results in the following test statistics :

$$t_a = \overline{\zeta}_a^{\,T}\, \mathbf{F}_{aa}^{-1}\, \overline{\zeta}_a \ , \tag{17}$$

where $\overline{\zeta}_a$ is the partial residual (score). If $t_a \geq t_b$, the component responsible for the fault is considered to be a rather than b.

Another statistical solution to the problem of isolating η_a consists in viewing parameter η_b as a nuisance, and using an existing method for inferring part of the parameters while ignoring and being robust to the complementary part. This method is called *min-max approach*. It consists in replacing the nuisance parameter component η_b by its least favorable value, for deciding between $\eta_a = 0$ and $\eta_a \neq 0$, with η_b unknown. This results in the following test statistics :

$$t_a^* = \overline{\zeta}_a^{*\,T}\, \mathbf{F}_a^{*-1}\, \overline{\zeta}_a^{\,*} \;, \tag{18}$$

where $\overline{\zeta}_a^{\,*} \triangleq \overline{\zeta}_a - \mathbf{F}_{ab}\, \mathbf{F}_{bb}^{-1}\, \overline{\zeta}_b$ is the effective residual (score) resulting from the regression of the informative partial score $\overline{\zeta}_a$ over the nuisance partial score $\overline{\zeta}_b$, and where the Schur complement $\mathbf{F}_a^* = \mathbf{F}_{aa} - \mathbf{F}_{ab}\, \mathbf{F}_{bb}^{-1}\, \mathbf{F}_{ba}$ is the associated Fisher information matrix. If $t_a^* \geq t_b^*$, the component responsible for the fault is considered to be a rather than b.

The properties and relationships of these two types of tests are investigated in [28].

3.4.2. Diagnostics.

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 in module 4.2 , 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 (6) for the components of the parameter vector Φ. Typically this results in the following type of directional test :

$$\chi^2_\Phi = \zeta^T\, \Sigma^{-1}\, \mathcal{J}\, \mathcal{J}_{\Phi\theta}\, \left(\mathcal{J}_{\Phi\theta}^T\, \mathcal{J}^T\, \Sigma^{-1}\, \mathcal{J}\, \mathcal{J}_{\Phi\theta}\right)^{-1} \mathcal{J}_{\Phi\theta}^T\, \mathcal{J}^T\, \Sigma^{-1}\, \zeta \gtrless \lambda \;. \tag{19}$$

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 described in module 4.2 , 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 described in module 4.2 , 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 :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & F\ X_k + V_{k+1} \\ Y_k & = & H\ X_k \end{array} \right. , \tag{20}$$

namely the $(\lambda, \varphi_\lambda)$ defined by :

$$\det\ (F - \lambda\ I) = 0, \quad (F - \lambda\ I)\ \phi_\lambda = 0, \quad \varphi_\lambda \triangleq H\ \phi_\lambda \tag{21}$$

The (canonical) parameter vector in that case is :

$$\theta \triangleq \left(\begin{array}{c} \Lambda \\ \mathrm{vec}\Phi \end{array} \right) \tag{22}$$

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 [39]. 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:

$$\mathcal{H}_{p+1,q} \triangleq \left(\begin{array}{cccc} R_0 & R_1 & \vdots & R_{q-1} \\ R_1 & R_2 & \vdots & R_q \\ \vdots & \vdots & \vdots & \vdots \\ R_p & R_{p+1} & \vdots & R_{p+q-1} \end{array} \right) \triangleq \mathrm{Hank}\left(R_i\right) \tag{23}$$

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 (9) lead to the well known key factorizations :

$$\begin{array}{rcl} R_i & = & HF^iG \\ \mathcal{H}_{p+1,q} & = & \mathcal{O}_{p+1}(H,F)\ \mathcal{C}_q(F,G) \end{array} \tag{24}$$

where:

$$\mathcal{O}_{p+1}(H,F) \triangleq \left(\begin{array}{c} H \\ HF \\ \vdots \\ HF^p \end{array} \right) \quad \text{and} \quad \mathcal{C}_q(F,G) \triangleq (G\ FG\ \cdots\ F^{q-1}G) \tag{25}$$

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 (10).

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 (9) yields the following representation of the observability matrix:

$$\mathcal{O}_{p+1}(\theta) = \begin{pmatrix} \Phi \\ \Phi\Delta \\ \vdots \\ \Phi\Delta^p \end{pmatrix} \tag{26}$$

where $\Delta \triangleq \text{diag}(\Lambda)$, and Λ and Φ are as in (11). Whether a nominal parameter θ_0 fits a given output covariance sequence $(R_j)_j$ is characterized by [1]:

$$\mathcal{O}_{p+1}(\theta_0) \text{ and } \mathcal{H}_{p+1,q} \text{ have the same left kernel space.} \tag{27}$$

This property can be checked as follows. From the nominal θ_0, compute $\mathcal{O}_{p+1}(\theta_0)$ using (15), and perform e.g. a singular value decomposition (SVD) of $\mathcal{O}_{p+1}(\theta_0)$ for extracting a matrix U such that:

$$U^T\, U = I_s \text{ and } U^T\, \mathcal{O}_{p+1}(\theta_0) = 0 \tag{28}$$

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:

$$U(\theta_0)^T\, \mathcal{H}_{p+1,q} = 0 \tag{29}$$

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 $\widehat{\mathcal{H}}_{p+1,q}$:

$$\widehat{\mathcal{H}}_{p+1,q} \triangleq \text{Hank}\left(\widehat{R}_i\right), \quad \widehat{R}_i \triangleq 1/(N-i) \sum_{k=i+1}^{N} Y_k\, Y_{k-i}^T \tag{30}$$

and to define the residual vector:

$$\zeta_N(\theta_0) \triangleq \sqrt{N}\ \text{vec}\left(U(\theta_0)^T\ \widehat{\mathcal{H}}_{p+1,q}\right) \tag{31}$$

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 (18), 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 (11), and for deriving the residual. Factorization (13) 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].

IPSO Project-Team

3. Scientific Foundations

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

$$\begin{aligned} y'(t) &= f(y(t)), \\ y(0) &= y_0. \end{aligned} \tag{32}$$

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

$$\rho \circ \varphi_t = \varphi_t^{-1} \circ \rho = \varphi_{-t} \circ \rho. \tag{33}$$

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

$$\mathcal{M} = \{y \in \mathbb{R}^n; g(y) = 0\} \tag{34}$$

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 [56] 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:

$$\begin{array}{rcrcl} \dot{p}(t) & = & -\nabla_q H(p(t),q(t)) & \in & \mathbb{R}^d \\ \dot{q}(t) & = & \nabla_p H(p(t),q(t)) & \in & \mathbb{R}^d \end{array} \tag{35}$$

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

$$\begin{array}{rcl} \dot{y}(t) & = & f(y(t),z(t)), \\ 0 & = & g(y(t)), \end{array} \tag{36}$$

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

$$\ddot{q} \quad = \quad -\nabla V(q) \tag{37}$$

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:

$$i\dot{\psi}(t) = \frac{1}{\varepsilon} H(t)\psi(t), \tag{38}$$

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

$$i\varepsilon \frac{\partial \psi}{\partial t} = H\psi, \tag{39}$$

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 [60], [59] 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:

$$\psi_1 = \exp\left(-i(\delta t)V/2\right) \exp\left(i(\delta t)\Delta\right) \exp\left(-i(\delta t)V/2\right)\psi_0 \tag{40}$$

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

$$i\alpha_\varepsilon u_\varepsilon(x) + \varepsilon^2 \Delta_x u_\varepsilon + n^2(x) u_\varepsilon = f_\varepsilon(x). \tag{41}$$

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,

$$i\partial_t \psi(t,x) = -\frac{1}{2}\Delta_x \psi + V(x)\psi, \tag{42}$$

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 :

$$\partial_t f(t,x,v) = \int_{\mathbf{R}^3} \sigma(v,v') \, [f(t,x,v') - f(t,x,v)] dv'. \tag{43}$$

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.

MATHRISK Team (section vide)

MAXPLUS Project-Team

3. Scientific Foundations

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 [102]. 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 [95], [122], [173], [178], [179]. Les semi-anneaux de type max-plus sont bien sûr reliés aux algèbres de Boole [82]. L'algèbre max-plus apparaît de manière naturelle en contrôle optimal et dans la théorie des équations aux dérivées partielles d'Hamilton-Jacobi [161], [160], [144], [128], [119], [165], [138], [120], [105], [66]. Elle apparaît aussi en analyse asymptotique (asymptotiques de type WKB [143], [144], [128], grandes déviations [159], asymptotiques à température nulle en physique statistique [84]), 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 [129], [149], [141], [73], laquelle a de nombreuse motivations, telles l'économie mathématique [146], et la théorie des jeux [162], [56]. 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 [83], [133], 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 [168], [125], [169], [130], [151]. Notons finalement, pour information, que l'algèbre max-plus est apparue récemment en géométrie algébrique [118], [172], [145], [171] et en théorie des représentations [107], [76], 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 [102]. 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 [95], [122], [173], [178], [179]. Of course, max-plus type semirings are related to Boolean algebras [82]. Max-plus algebras arises naturally in optimal control and in the theory of Hamilton-Jacobi partial differential equations [161], [160], [144], [128], [119], [165], [138], [120], [105], [66]. It arises in asymptotic analysis (WKB asymptotics [143], [144], [128], large deviation asymptotics [159], or zero temperature asymptotics in statistical physics [84]), 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 [129], [149], [141], [73], a theory with numerous motivations, including mathematical economy [146] and game theory [162], [56]. 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 [83], [133] 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 [168], [125], [169], [130], [151]. We finally note for information that max-plus algebra has recently arisen in algebraic geometry [118], [172], [145], [171] and in representation theory [107], [76], 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:

$$v_{ij}(k) = \min_{\ell:\ \ell_0=i,\ \ell_k=j} \sum_{r=0}^{k-1} M_{\ell_r \ell_{r+1}} \ , \tag{44}$$

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:

$$v_{ij}(k) = \min_{1 \le s \le n} (M_{is} + v_{sj}(k-1)) \ . \tag{45}$$

On reconnaît ainsi une équation linéaire min-plus :

$$v(k) = Mv(k-1) \ , \tag{46}$$

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,

$$v(x,T) = \inf_{X(\cdot),\ X(0)=x} \int_0^T L(X(t), \dot{X}(t)) \mathrm{d}t + \phi(X(T)) \ , \tag{47}$$

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,

$$v(\cdot, 0) = \phi, \quad \frac{\partial v}{\partial T} + H(x, \frac{\partial v}{\partial x}) = 0, \qquad H(x,p) = \sup_{y \in \mathbb{R}^n} \left(-p \cdot y - L(x,y)\right) \ , \tag{48}$$

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 [144], [128], [138]).

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 [161], [173], [95], [121], [89], [72], [6] pour la dimension finie, et [144], [128] 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" [105].

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 [144], [128], [138]).

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 [161], [173], [121], [89], [95], [72], [6] for the finite dimension case, and [144], [128] 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 [105].

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 :

$$\begin{array}{lrcl} \text{monotonie}/\textit{monotonicity} & x \leq y & \Rightarrow & f(x) \leq f(y) \ , \\ \text{contraction}/\textit{nonexpansiveness} & \|f(x) - f(y)\|_\infty & \leq & \|x - y\|_\infty \ . \end{array} \tag{49}$$

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 [94] (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 [56]. 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 [149] pour un panorama. Signalons que dans [111],[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 le §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 [94] (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 [56]. 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 [149] for more background. In [111],[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* [160], [97], [96], [66],[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 [143]. 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 [131], [63].

English version

Another point of view on optimal control is the theory of *Bellman processes* [160], [97], [96], [66], [1] which provides a max-plus analogue of probability theory, relying on the theory of *idempotent measures* due to Maslov [143]. 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 [131], [63].

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 [85], 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 [89] 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 [150], [124], ont aussi été étudiées. Les systèmes dynamiques max-plus linéaires aléatoires sont particulièrement utiles dans la modélisation des réseaux [71]. Les modèles d'automates à multiplicités max-plus [109], incluant certains versions temporisées des modèles de traces ou de tas de pièces [113], 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 [168], [125], [137], [169], [130], [151]. Ils fournissent des modèles particulièrement adaptés à l'analyse de problèmes d'ordonnancement [136].

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 [85], we get min-plus linear equations similar to (2). This observation, or rather, the dual observation concerning dater functions, was the starting point [89] 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 [150], [124], have been particularly studied. Random max-plus linear dynamical systems are particularly useful in the modelling of networks [71]. Max-plus automata models [109], which include some timed version of trace or heaps of pieces models [113], allow to represent phenomena of concurrency or resource sharing. Note that max-plus automata have been much studied in theoretical computer science [168], [125], [137], [169], [130], [151]. Such automata models are particularly adapted to the analysis of scheduling problems [136].

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 [102], [78], [173], [178],[6]. Les versions dimension infinie du problème $Ax = b$ sont reliées aux questions d'analyse convexe abstraite [170], [163], [61] 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 [123], [152],[6]). Le sujet fait l'objet d'un intérêt récemment renouvelé [98].

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 [102], [78], [173], [178],[6]. Infinite dimensional versions of the $Ax = b$ problem are related to questions of abstract convex analysis [170], [163], [61] and nonconvex duality. 3) The general linear system $Ax = Bx$ leads to interesting combinatorial developments (max-plus polyedra, determinants, symmetrisation [123], [152],[6]). The subject has attracted recently a new attention [98].

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

$$e^{-a/\epsilon} + e^{-b/\epsilon} \asymp e^{-\min(a,b)/\epsilon} \ , \qquad e^{-a/\epsilon} \times e^{-b/\epsilon} = e^{-(a+b)/\epsilon} \ , \tag{50}$$

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 [143], 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], [159], [62], pour les relations entre l'algèbre max-plus et les grandes déviations, voir aussi [60], [59], [58] 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 [172].

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 [143], 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], [159], [62] for the relation between max-plus algebra and large deviations. See also [60], [59], [58] 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 [172].

MC2 Project-Team

3. Scientific Foundations

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 $50\mu m \times 50\mu m$). 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 [71]. 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. 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 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: Angelo Iollo, Thierry Colin, 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 [81] where the efficacy of radiotherapy is studied and in [82] 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 [50], [72], [73]. 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 [81].

A forthcoming investigation in cancer treatment simulation is the influence of the electrochemotherapy [76] 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 [80] 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 [59]. 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 [51]. 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 [66]. 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) [74]. 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-Witney 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 [55]. 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 [56]. 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 [70].

A similar procedure is also used in the estimation based on gappy POD, see [85] and [89]. 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 [62] 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 [75] 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 [84]. 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 Team

3. Scientific Foundations

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 the physics; the equations governing their behavior is then rather 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 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: [56], [42]– 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 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 adapted 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 of 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 [48], [63].

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 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 [1].

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 [24], [60], [34], [72]. 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, 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, adapted 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 [34], [37]. On the other hand the *value function* –the minimum of the criteria at each point– would give the solution to both questions if it were differentiable, but it is usually not, and is only a viscosity solution of the Hamilton-Jacobi-Bellman (**HJB**) equation (a PDE), whose study requires sophisticated non-smooth and singularity analysis.

[1] 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" [51] 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" [27]. 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, 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. [43], [52], [53], [67], [68], [70], [71], [44] insist on the construction of continuous feedback, hence smooth CLFs whereas, on the contrary, [41], [74], [75], [76], [77] 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

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 [65] and Kantorovitch [57] to the recent revival initiated by fundamental contributions due to Brenier [38], [39] and McCann [64]. However, the study of the same transportation problems in the presence of non-holonomic constraints –(like being an admissible trajectory for a control system– is quite new. The first contributors were Ambrosio and Rigot [25] 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 [22], then by Figalli and Rifford [46] 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 [61] and Sturm [79], [80]. A pioneer work has been work in the Heisenberg group by Juillet [54] who captured the right notion of curvature in this setting. Agrachev and Lee [23] 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.

One of the results of A. Hindawi's PhD under the supervision of L. Rifford and J.-B. Pomet was to extend regularity theory established in the Euclidean case to the more general quadratic costs associated with linear optimal control problems (LQR), see [50]. This successful result opens a new range of optimal transport problems associated with cost coming from optimal control problems. We can nowadays expect regularity properties for optimal transport maps associated with reasonable optimal control problems with constraints on the state or on the velocities.

We believe that matching optimal transport with geometric control theory is one originality of our team. We expect interactions in both ways.

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 [26], [78].

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 [30], [31], [40], [49]. Without deciding the control a priori, an "average control system" may be defined as in [2].

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 lime 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 [31] 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.

MICMAC Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

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 minerals have 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 project-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.

MISTIS Project-Team

3. Scientific Foundations

3.1. Mixture models

Participants: Angelika Studeny, Thomas Vincent, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Jean-Baptiste Durand, Florence Forbes, Stéphane Girard, Marie-José Martinez, Darren Wraith.

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,

$$f(y_i \mid \theta) = \sum_{k=1}^{K} P(z_i = k \mid \theta) f(y_i \mid z_i, \theta) \ . \tag{51}$$

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 are 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, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Jean-Baptiste Durand, Florence Forbes, Darren Wraith.

graphical models, Markov properties, conditional independence, hidden Markov trees, clustering, statistical learning, missing data, mixture of distributions, EM algorithm, stochastic algorithms, selection and combination of models, statistical pattern recognition, image analysis, hidden Markov field, Bayesian inference

Graphical modelling provides a diagrammatic representation of the logical 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 the mean field principle 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: El-Hadji Deme, Jonathan El-Methni, Ludovic Leau-Mercier, Stéphane Girard, Gildas Mazo, Kai Qin, Huu Giao Nguyen, Farida Enikeeva, Seydou-Nourou Sylla.

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 [59]. 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 [68] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [58], 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).

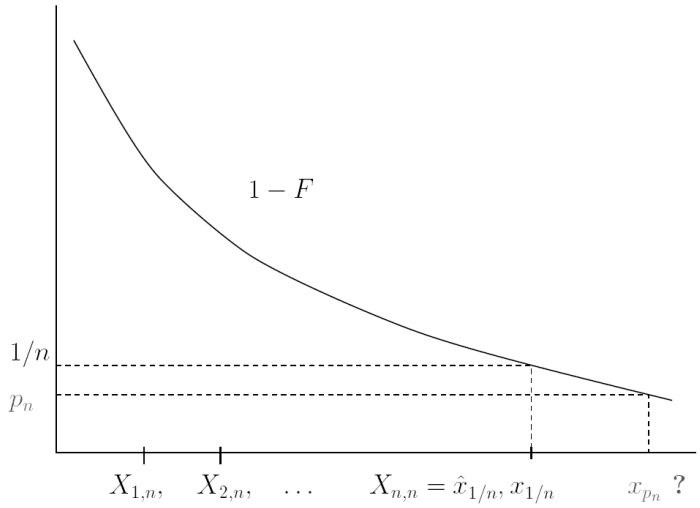


Figure 1. The curve represents the survival function $x \to P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\widehat{x}_{1/n} = X_{n,n}$. As illustrated in the figure, to estimate p_n-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.

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:

$$P(X > x) = x^{-1/\theta}\ell(x),\ x > x_0 > 0, \tag{52}$$

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

$$\frac{\ell(tx)}{\ell x} \to 1 \text{ as } x \to \infty \tag{53}$$

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:

$$P(X > x) = \exp\{-x^{\theta}\ell(x)\},\ x > x_0 > 0. \tag{54}$$

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 [62]. 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 [53]. 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 [68].

MODAL Project-Team

3. Scientific Foundations

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) space, 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, 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.

NACHOS Project-Team

3. Scientific Foundations

3.1. High order discretization methods

The applications in computational electromagnetics and computational geoseismics that are considered by the team lead to the numerical simulation of wave propagation in heterogeneous media or/and involve irregularly shaped objects or domains. 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 overall objective of the research activities undertaken by the team is to develop numerical methods putting the emphasis on several features:

- Accuracy. The foreseen numerical methods should ideally rely on discretization techniques that best fit to the geometrical characteristics of the problems at hand. For this reason, the team focuses on methods working on unstructured, locally refined, even non-conforming, simplicial meshes. 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, two main 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.

- Numerical efficiency. The simulation of unsteady problems most often relies on explicit time integration schemes. Such schemes are constrained by a stability criteria linking the 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. In order to improve this situation, one possible approach consists in resorting to an implicit time scheme in regions of the computational domain where the underlying mesh is refined while an explicit time scheme is applied to the remaining part of the 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, more classical approach is to devise a local time strategy in the context of a fully explicit time integration scheme. Stability and accuracy are still important challenges in this case.

  On the other hand, 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 lead to 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). Besides, appropriate parallelization strategies need to be designed that combine SIMD and MIMD programming paradigms. Moreover, maximizing the effective floating point performances will require the design of numerical algorithms that can benefit from the optimized BLAS linear algebra kernels.

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 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 non-linear convection. The success of the finite volume method is due to its ability to capture discontinuous solutions which may occur when solving non-linear equations or more simply, when convecting discontinuous initial data in the linear case. Let us first remark that DG methods share with finite volumes this property since a first order finite volume scheme can be viewed as a 0th order DG scheme. However a DG method may be also considered as a finite element one where the continuity constraint at an element interface is released. While it keeps almost all the advantages of the finite element 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 [43]- [44]- [45]- [48]:

- 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 discretization in space is coupled to an explicit time integration method, the DG method leads to a block diagonal mass matrix independently of the form of the local approximation (e.g the type of polynomial interpolation). This is a striking difference with classical, continuous finite element formulations. Moreover, the mass matrix is diagonal if an orthogonal basis is chosen.
- It easily handles complex meshes. The grid may be a classical conforming finite element 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 to the design of a hp-adaptive solution strategy (i.e where the characteristic mesh size h and the interpolation degree p changes locally wherever it is needed).
- It is 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 finite element 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 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 finite volume 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.

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 [14]-[3]. This DG method combines a central numerical flux function for the approximation of the integral term at an interface between 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 [12]-[13], 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 [5]-[11]. For the numerical treatment of the time harmonic Maxwell equations, we have studied similar DG methods [7]-[18] and more recently, HDG (Hybridized Discontinuous Galerkin) methods [22].

3.2. Domain decomposition methods

Domain Decomposition (DD) methods are flexible and powerful techniques for the parallel numerical solution of systems of PDEs. As clearly described in [51], they can be used as a process of distributing a computational domain among a set of interconnected processors or, for the coupling of different physical models applied in different regions of a computational domain (together with the numerical methods best adapted to each model) and, finally as a process of subdividing the solution of a large linear system resulting from the discretization of a system of PDEs into smaller problems whose solutions can be used to devise a parallel preconditioner or a parallel solver. In all cases, DD methods (1) rely on a partitioning of the computational domain into subdomains, (2) solve in parallel the local problems using a direct or iterative solver and, (3) call for an iterative procedure to collect the local solutions in order to get the global solution of the original problem. Subdomain solutions are connected by means of suitable transmission conditions at the artificial interfaces between the subdomains. The choice of these transmission conditions greatly influences the convergence rate of the DD method. One generally distinguish three kinds of DD methods:

- Overlapping methods use a decomposition of the computational domain in overlapping pieces. The so-called Schwarz method belongs to this class. Schwarz initially introduced this method for proving the existence of a solution to a Poisson problem. In the Schwarz method applied to the numerical resolution of elliptic PDEs, the transmission conditions at artificial subdomain boundaries are simple Dirichlet conditions. Depending on the way the solution procedure is performed, the iterative process is called a Schwarz multiplicative method (the subdomains are treated sequentially) or an additive method (the subdomains are treated in parallel).

- Non-overlapping methods are variants of the original Schwarz DD methods with no overlap between neighboring subdomains. In order to ensure convergence of the iterative process in this case, the transmission conditions are not trivial and are generally obtained through a detailed inspection of the mathematical properties of the underlying PDE or system of PDEs.

- Substructuring methods rely on a non-overlapping partition of the computational domain. They assume a separation of the problem unknowns in purely internal unknowns and interface ones. Then, the internal unknowns are eliminated thanks to a Schur complement technique yielding to the formulation of a problem of smaller size whose iterative resolution is generally easier. Nevertheless, each iteration of the interface solver requires the realization of a matrix/vector product with the Schur complement operator which in turn amounts to the concurrent solution of local subproblems.

Schwarz algorithms have enjoyed a second youth over the last decades, as parallel computers became more and more powerful and available. Fundamental convergence results for the classical Schwarz methods were derived for many partial differential equations, and can now be found in several books [51]- [50]- [53].

The research activities of the team on this topic aim at the formulation, analysis and evaluation of Schwarz type domain decomposition methods in conjunction with discontinuous Galerkin approximation methods on unstructured simplicial meshes for the solution of time domain and time harmonic wave propagation problems. Ongoing works in this direction are concerned with the design of non-overlapping Schwarz algorithms for the solution of the time harmonic Maxwell equations. A first achievement has been a Schwarz algorithm for the time harmonic Maxwell equations, where a first order absorbing condition is imposed at the interfaces between neighboring subdomains [9]. This interface condition is equivalent to a Dirichlet condition for characteristic variables associated to incoming waves. For this reason, it is often referred as a natural interface condition. Beside Schwarz algorithms based on natural interface conditions, the team also investigates algorithms that make use of more effective transmission conditions [10]. Recent contributions are concerned with the design and anlysis of such optimized Schwarz algorithm for the solution of the time harmonic Maxwell equations with non-zero conductivity [17].

3.3. 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 Team

3. Scientific Foundations

3.1. Overview

The adaptive simulation algorithms we develop typically consist in two main components. The first one determines *which degrees of freedom are simulated* at a give time step, based on the current system's state, as well as user-defined precision or cost thresholds. The second component *incrementally updates the system's state* based on the set of active degrees of freedom. In particular, incremental algorithms update the system's potential energy and forces. This allows the user to smoothly trade between precision and cost.

We detail this approach in two important types of simulations: Cartesian quasi-statics and torsion-angle dynamics. A novel, very general approach for adaptive dynamics simulations of particles — that has a number of important benefits over previous approaches — is mentioned in more detail in Section 6.1 .

3.2. Adaptive Cartesian mechanics

In order to focus computations on a specific set of atoms, when performing quasi-static simulations (minimizations), we have developed an adaptive Cartesian mechanics algorithm, which decides which atoms should move at each time step.

In the simplest approach, we simply examine the force applied on each atom. When the norm of the force is above a user-defined threshold, the atom is active. Else the atom position is frozen. A slightly more elaborate version consists in defining the threshold automatically based on the system state (it might be e.g., the average applied force, a percentage of the maximum norm, etc.).

In order to avoid the linear cost of determining the set of active atoms at each time step, a binary tree is used to represent the system. Each leaf node represents an individual atom, while each internal node represents a set of atoms. Each leaf node stores the norm of the force applied to the corresponding atom. Each non-leaf node stores the maximum of the two force norms of its children, as illustrated in Figure 2 . We use two tree passes in order to update tree nodes' values and to determine the new active atoms. In the first, bottom-up pass, force norms are updated in a sub-tree of the binary tree (only some atoms have moved since the previous time step, so only some forces have been updated), starting from the leaves with modified norms, in $O(k^{old}(log(\frac{n}{k^{old}})+1))$ times where k^{old} is the number of active atoms and n the total number of atoms. In the second, top-down pass, the new active atoms (i.e., the atoms with the force norms which are now the largest), are determined in $O(k^{new}(log(\frac{n}{k^{new}})+1))$ times where k^{new} is the new number of active atoms. This process is illustrated in Figure 2 as well.

Precisely, Figure 2 illustrates the procedure to determine the active zone, when the threshold is automatically set to half the largest atomic force norm. In this example, the four leaves correspond to atoms 1 to 4. The value indicated in each leaf node is the norm of the force applied to its corresponding atom. For internal nodes, this value is the maximum of the norms of the forces applied to atoms in the corresponding group. In step 0, the threshold is automatically set to 10. As a result, only atom 1 moves. In step 1, the potential is incrementally updated, and the norms of the forces applied to atoms 1 and 2 are updated. In step 2, the values associated to the tree nodes are incrementally updated through a bottom-up pass that starts from the modified leaf nodes values. Because of this bottom-up update, the adaptive threshold becomes equal to 4. In step 3, the new active atoms are determined through a top-down pass, by visiting only the nodes that have a value larger than the adaptive threshold.

3.3. Adaptive torsion-angle mechanics

In many situations, it is preferable to represent molecular systems as articulated bodies, and perform so-called *torsion-angle* mechanics. This may be to allow for larger time step sizes in a simulation, or because the user wants to focus to e.g., protein backbone deformations.

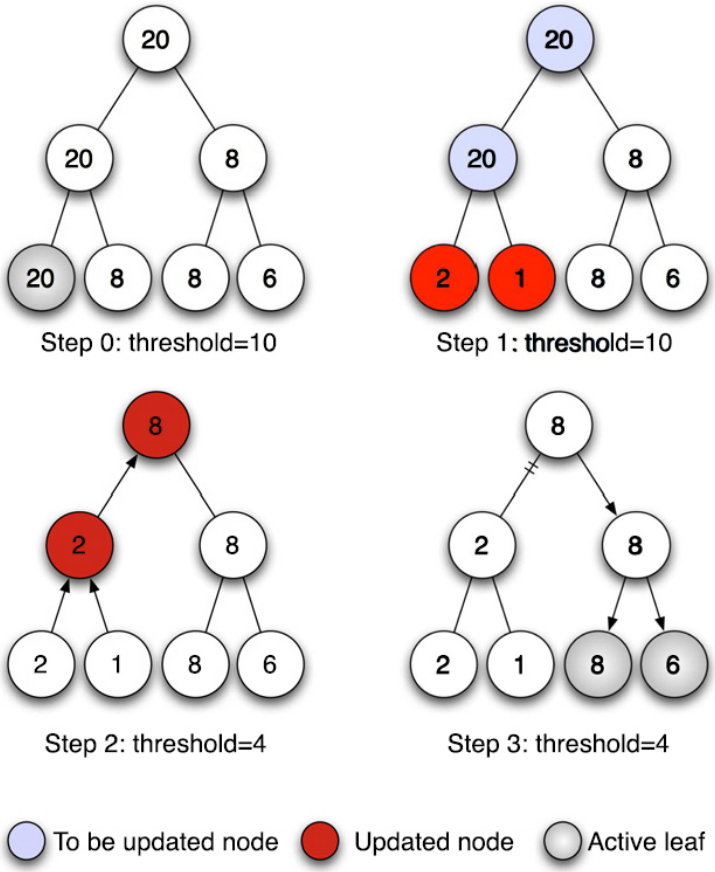


Figure 2. Adaptive Cartesian mechanics.

We have also developed an adaptive mechanics algorithm in the case of torsion-angle representations. In this case, a molecular system is recursively defined as the assembly of *two* molecular systems connected by a *joint* (when connecting two subassemblies which belong to the same molecule) or, more generally, by a *rigid body transform* (to assemble several molecules).

As in the Cartesian mechanics case, the complete molecular system is thus also represented by a binary tree, in which leaves are rigid bodies (a rigid body can be a single atom), internal nodes represent both sub-assemblies and connections between sub-assemblies, and the root represents the complete molecular system (see Figure 3 on the right, which shows an assembly tree associated to a short polyalanin). This hierarchical representation handles any branched molecule or groups of molecules, since the connections between two sub-molecular systems can be a rigid body transformation. In this representation, the positions of atoms are thus represented as superimposed rigid transformations: the position of any atom is obtained from the position of the whole set, to which is "added" the transformation from the complete set to the sub-set the atom belongs to, and so on until we reach the leaf node representing the atom. Similarly, the atomic motions are superimposed rigid motions.

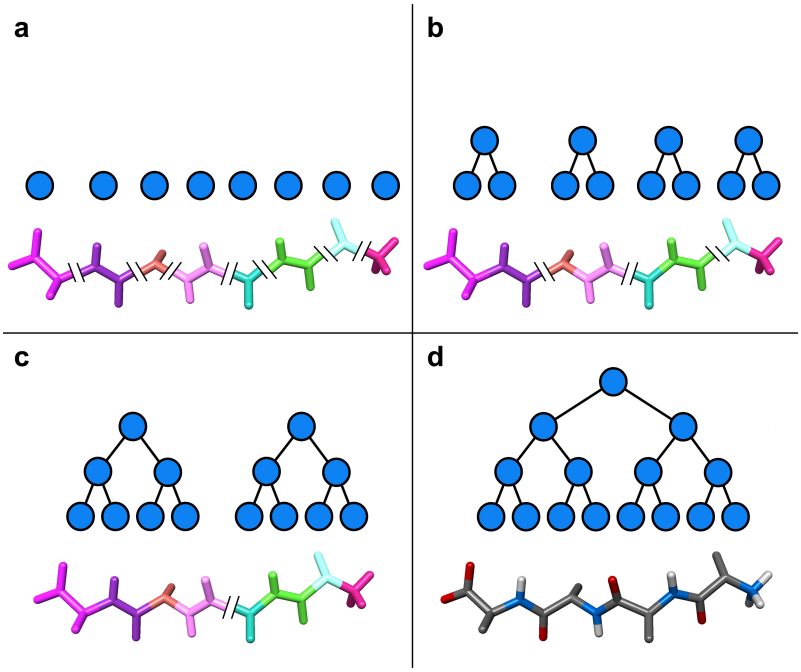


Figure 3. The assembly tree associated to a short polyalanin.

Our adaptive framework relies on two essential components. First, we associate a hierarchical set of reference frames to the assembly tree. Precisely, each node is associated to a local reference frame, in which all dynamical coefficients are expressed. This allows us to avoid updating these coefficients when a sub-assembly moves rigidly. Second, we have demonstrated that it is possible to determine a priori, at each time step, the set of joints which have the largest accelerations. Precisely, when going down the tree to compute joint accelerations, we are able to compute the weighted sum of the (squared) norms of joint accelerations in a sub-assembly C before computing joint accelerations themselves:

$$\mathcal{A}(C) = (\mathbf{f}^C)^T \Psi^C \mathbf{f}^C + (\mathbf{f}^C)^T \mathbf{p}^C + \eta^C, \tag{55}$$

where the right part is a quadratic form of the spatial forces applied on the "handles" of node C. This allows us to cull away those sub-assemblies with (relatively) lower internal accelerations, and focus on the most mobile joints. Thus, at each time step, we can thus predict the set of joints with highest accelerations without computing all accelerations, and we simulate only a sub-tree of the assembly tree (the green nodes in the assembly tree, as in the figure above), based on an user-defined error threshold or computation time constraints. This sub-tree is called the active region, and may change at each time step.

We have exploited these two characteristics - hierarchical coordinate systems and adaptive motion refinement - to develop data structures and algorithms which enable adaptive molecular mechanics. The key observation in our approach is the following: all coefficients which only depend on relative atomic positions do not have to be updated when these relative positions do not change. We can thus store in each node of the assembly tree partial system states which hold information relative only to the node itself.

Precisely, each time step involves the following operations:

1. Adaptive acceleration update
    1. Determination of the acceleration update region: we determine the acceleration update region, i.e., the subset of nodes of the full articulated body which matter the most according to the acceleration metric, as indicated above. The union of the previous active region and the acceleration update region is the transient active region, i.e., the region temporarily considered as the active region.
    2. Joint accelerations projection: the acceleration is projected on the reduced motion space defined by the transient active region (to ensure that joint accelerations are consistent with both motion constraints and applied forces).
2. Adaptive velocity update
    1. Determination of the new active region: we update the joint velocities and the velocity metric values of the nodes in the transient active region. We then determine the set of nodes which are considered to be important according to the velocity metric (which is similar to the acceleration metric). This set becomes the new active region.
    2. Joint velocities projection: if one or more nodes become inactive due to the update of the active region, we determine a set of impulses that we must apply to the transient hybrid body to perform the rigidification of these nodes. This amounts to projecting joint velocities to the reduced motion space defined by the new active region.
3. Adaptive position update
    1. Position update: we update joint positions based on non-zero joint velocities in the active region.
    2. State update: once joint positions have been updated, we update the rest of the system's state: inverse inertias, acceleration metric coefficients, partial neighbor lists, partial force tables, etc.

Again, each of these steps involves a limited sub-tree of the assembly tree, which enables a fine control of the compromise between computation time and precision.

We have showed that our adaptive approach allows for a number of applications, some of which that were not possible for classical methods when using low-end desktop workstations. Indeed, by selecting a sufficiently small number of simultaneously active degrees of freedom, it becomes possible to perform interactive structural modifications of complex molecular systems.

NECS Project-Team

3. Scientific Foundations

3.1. Multi-disciplinary nature of the project

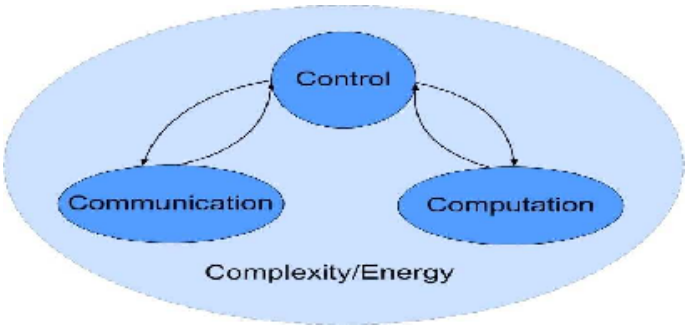


Figure 2. Relation of the NCS area with the fields of: Control, Communication, Computation.

The team's project is to investigate problems in the area of NCS with the originality of integrated aspects on computation, communication and control. The combination of these three disciplines requires the interplay of the multi-disciplinary fields of: communication, real-time computation, and systems theory (control). Figure 2 , shows the natural interaction between disciplines that concern the NeCS project. The arrows describe the direction in which these areas interact, i.e.

(a) *Control in Communication*
(b) *Communication in Control*
(c) *Computation in Control*
(d) *Control in Computation*

Complexity and energy-management are additional features to be considered as well. Complexity here refers to the problems coming from: wireless networks with varying interconnection topologies, multi-agent systems coordination, scaling with respect to a growing number of sensors. Energy management concerns in particular the efficient handling of energy in wireless sensors, and means an efficient way to handle both information transmission and computation.

3.1.1. (a) Control in Communication

This topic is the study of how control-theoretic methods can be applied in order to solve some problems found in the communication field. Examples are: the Power control in cell telephones, and the optimal routing of messages in communication networks (Internet, sensor networks).

3.1.2. (b) Communication in Control

This area concerns problems where communication and information theory interact with systems theory (control). As an example of a classical paradigm we can mention the stabilization problem under channel (communication) constraints. A key result here [75] was to show that it was generically impossible to stabilize a linear system in any reasonable sense, if the feedback channel's Shannon classical capacity C was smaller than the sum of the logarithms, base 2, of the unstable eigenvalues. In other words, in order to be able to cope with the stabilization problem under communication constraints, we need that

$$C > \sum_i \log_2 \lambda_i$$

where the $\lambda_i' s$ are unstable eigenvalues of the open loop system. Intuitively, this means that the rate of information production (for discrete-time linear systems, the intrinsic rate bits/time equals $\sum_i \log_2 \lambda_i$) should be smaller than the rate of information that can be transmitted throughout the channel. In that way, a potentially growing signal can be cached out, if the information of the signal is sent via a channel with fast enough transmission rate. In relation to this, a problem of interest is the coding and control co-design. This issue is motivated by applications calling for data-compression algorithms aiming at reducing the amount of information that may be transmitted throughout the communication channel, and therefore allowing for a better resource allocation and/or for an improvement of the permissible closed loop system bandwidth (data-rate). Networked controlled systems also constitute a new class of control systems including specific problems concerned by delays. In NCS, the communication between two agents leads unavoidably to transmission delays. Also, transmission usually happens in discrete time, whereas most controlled processes evolve in continuous time. Moreover, communication can induce loss of information. Our objectives concern the stabilization of systems where the sensor, actuator and system are assumed to be remotely commissioned by a controller that interchanges measurements and control signals through a communication network. Additional dynamics are introduced due to time-varying communication delays, asynchronous samplings, packets losses or lack of synchronization. All those phenomena can be modeled as the introduction of time-delays in the closed loop system. Even if these time-delay approaches can be easily proposed, they require careful attention and more complex analysis. In general, the introduction of delays in a controlled loop leads to a reduction of the performance with respect to the delay-free situation and could even make the systems unstable. Our objective is to provide specific modeling of these phenomena and to develop dedicated tools and methodologies to cope with stability and stabilization of such systems.

3.1.3. (c) Computation in Control

This area concerns the problem of redesigning the control law such as to account for variations due to the resource allocation constraints. Computation tasks having different levels of priority may be handled by asynchronous time executions. Hence controllers need to be re-designed as to account for non-uniform sampling times resulting in this framework. Questions on how to redesign the control laws while preserving its stability properties are in order. This category of problems can arise in embedded systems with low computation capacity or low level resolution.

3.1.4. (d) Control in Computation

The use of control methods to solve or to optimize the use of computational resources is the key problem in this area. This problem is also known as a scheduling control. The resource allocations are decided by the controller that tries to regulate the total computation load to a prefixed value. Here, the system to be regulated is the process that generates and uses the resources, and not any physical system. Hence, internal states are computational tasks, the control signal is the resource allocation, and the output is the period allowed to each task.

3.1.5. (c + d) Integrated control/scheduling co-design

Control and Computation co-design describes the possibility to study the interaction or coupling between the flows (c) and (d). It is possible, as shown in Fig. 3 , to re-frame both problems as a single one, or to interpret such an interconnection as the cascade connection between a computational system, and a physical system. In our framework the feedback scheduling is designed w.r.t. a QoC (Quality of Control) measure. The QoC criterion captures the control performance requirements, and the problem can be stated as QoC optimization under constraint of available computing resources. However, preliminary studies suggest that a direct synthesis of the scheduling regulator as an optimal control problem leads, when it is tractable, to a solution too costly to be implemented in real-time applications [64]. Practical solutions will be found in the currently available control theory and tools or in enhancements and adaptation of current control theory. We propose in Fig. 3

a hierarchical control structure: besides the usual process control loops we add an outer control loop whose goal is to manage the execution of the real-time application through the control of the scheduling parameters of the inner loops. Together with the outer loop (working on a periodic sampled time scale) we also need a scheduling manager working on a discrete events time scale to process exception handling and admission control. The task periods directly affect the computing load, they have been chosen as actuators. They can be implemented through software variable clocks. As timing uncertainties cannot be avoided and are difficult to model or measure, we currently design robust control algorithms using the H_∞ control theory, which have been successfully simulated and experimentally validated [74]. This methodology is supported by the software ORCCAD (see Section 5.1) where a run-time library for multi-rate multitasking has been developed and integrated. It will be further improved using a QoS-based management of the timing constraints to fully benefit from the intrinsic robustness of closed loop controllers w.r.t. timing uncertainties.

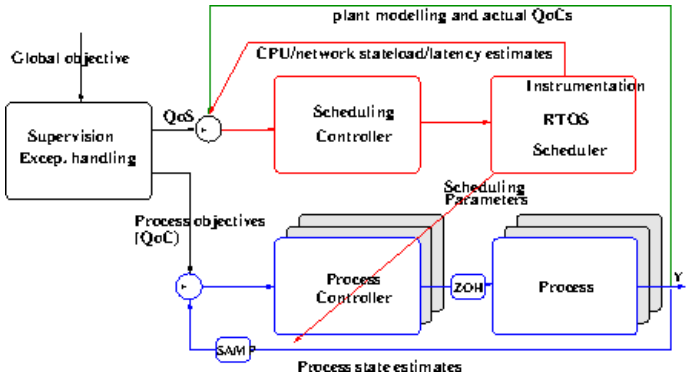


Figure 3. Hierarchical control structure.

3.2. Main Research Directions

The main objective of the project is to develop a unified control, communication, computing co-design methodology explicitly accounting for all the components involved in the system controlled over a network. This includes quantifier properties, scheduling parameters, encoder/decoder, alphabet length, bandwidth of the transmission media (wire or wireless), delays, resource allocation, jitter, etc. These components, including the control laws, should be designed so as to optimize performance/stability trade-offs resulting from the ceiling of the computing resources, the channel capacity limitations and the quality of the send/received information protocols. More informations about the main research directions of the team can be found in [1], [3],[2], [4] [5], [6], [7], [8], [9] and [10].

In short, the project is centered along the following 3 main axes:

1. **Control under Communications Constraints.** One well established topic along this axis concerns the coding and control co-design. That is, the design of new code alphabets simultaneously than the design of the control law. Or equivalently, the ability of designing codes containing information pertained to the system model and the control law. The objective being the improvements of the overall closed loop performances. Besides this matter, additional improvements pertain to the field of the information theory are also in order.
2. **Control under Computational resources constraints.** The main objective here is the design of control loops by explicitly accounting for the network and/or the computing resources. Dynamic allocation of such resources depends on the desired controlled systems specifications. Keys aspects to be considered are: the design of controllers with variable sampling time, the robustness with

respect to time uncertainties such as the input/output latencies, the global control of resources and its impact over the performance and the robustness of the system to be controlled. We aim to provide an integrated control and scheduling co-design approach [1].

3. **Controlling Complexity.** Design and control of partially cooperative networked (possible also multi-agent) systems subject to communication and computational constraints. Here, a large number of entities (agents), having each its own goal share limited common resources. In this context, if there is no minimum coordination, dramatic consequences may follow, on the other hand, total coordination would be impossible because of the lack of exhaustive, reliable and synchronous information. Finally, a local network of strategies that are based on worst-case assumptions is clearly far from being realistic for a well designed system. The aim of this topic is to properly define key concepts and the relevant variables associated to the above problem (sub-system, partial objective, constraints on the exchanged data and computational resources, level of locally shared knowledge, key parameters for the central level, etc).

NON-A Project-Team

3. Scientific Foundations

3.1. Fast parametric estimation and its applications

Parametric estimation may often be formalized as follows:

$$y = F(x, \Theta) + n, \tag{56}$$

where:

- the measured signal y is a functional F of the "true" signal x, which depends on a set Θ of parameters,
- n is a noise corrupting the observation.

Finding a "good" approximation of the components of Θ has been the subject of a huge literature in various fields of applied mathematics. Most of those researches have been done in a probabilistic setting, which necessitates a good knowledge of the statistical properties of n. Our project is devoted to a new standpoint, which does not require this knowledge and which is based on the following tools, which are of algebraic flavor:

- differential algebra [2], which plays with respect to differential equations a similar role that the commutative algebra plays with respect to algebraic equations;
- module theory, i.e. linear algebra over rings, which are not necessarily commutative;
- operational calculus, which is the most classical tool among control and mechanical engineers [3].

3.1.1. Linear identifiability

In the most problems, which appear in linear control as well as in signal processing, the unknown parameters are *linearly identifiable*: standard elimination procedures are yielding the following matrix equation

$$P \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_r \end{pmatrix} = Q, \tag{57}$$

where:

- $\theta_i, 1 \leq i \leq r$ represents unknown parameter,
- P is a $r \times r$ square matrix and Q is a $r \times 1$ column matrix,
- the entries of P and Q are finite linear combinations of terms of the form $t^\nu \frac{d^\mu \xi}{dt^\mu}$, $\mu, \nu \geq 0$, where ξ is an input or output signal,
- the matrix P is *generically* invertible, i.e., $\det(P) \neq 0$.

3.1.2. How to deal with perturbations and noises?

With noisy measurements equation (2) becomes:

[2] Differential algebra was introduced in nonlinear control theory by one of us almost twenty years ago for understanding some specific questions like input-output inversion. It allowed us to recast the whole of nonlinear control into a more realistic light. The best example is of course the discovery of *flat* systems, which are now quite popular in industry.

[3] Operational calculus is often formalized *via* the Laplace transform whereas the Fourier transform is today the cornerstone in estimation. Note that the one-sided Laplace transform is causal, but the Fourier transform over R is not.

$$P \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_r \end{pmatrix} = Q + R, \tag{58}$$

where R is a $r \times 1$ column matrix, whose entries are finite linear combinations of terms of the form $t^\nu \frac{d^\mu \eta}{dt^\mu}$, $\mu, \nu \geq 0$, where η is a perturbation or a noise.

3.1.2.1. Structured perturbations

A perturbation π is said to be *structured* if, and only if, it can be annihilated by a linear differential operator of the form $\sum_{\text{finite}} a_k(t) \frac{d^k}{dt^k}$, where $a_k(t)$ is a rational function of t, i.e. $\left(\sum_{\text{finite}} a_k(t) \frac{d^k}{dt^k}\right) \pi = 0$. Note that many classical perturbations, like a constant bias, are annihilated by such an operator. An *unstructured* noise cannot be annihilated by a non-zero differential operator.

By well-known properties of the non-commutative ring of differential operators, we can multiply both sides of equation (3) by a suitable differential operator Δ such that equation (3) becomes:

$$\Delta P \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_r \end{pmatrix} = \Delta Q + R', \tag{59}$$

where the entries of the $r \times 1$ column matrix R' are unstructured noises.

3.1.2.2. Attenuating unstructured noises

Unstructured noises are usually dealt with stochastic processes like white Gaussian noises. They are considered here as highly fluctuating phenomena, which may therefore be attenuated *via* low pass filters. Note that no precise knowledge of the statistical properties of the noises is required.

3.1.2.3. Comments

Although the previous noise attenuation procedure [4] may be fully explained *via* formula (4), its theoretical comparison [5] with today's literature [6] has yet to be done. It will require a complete resetting of the notions of noises and perturbations. Besides some connections with physics, it might lead to quite new "epistemological" issues [101].

3.1.3. Some hints on the calculations

The time derivatives of the input and output signals appearing in equations (2), (3), (4) can be suppressed in the two following ways which might be combined:

- integrate both sides of the equation a sufficient number of times,
- take the convolution product of both sides by a suitable low pass filter.

The numerical values of the unknown parameters $\Theta = (\theta_1, \cdots, \theta_r)$ can be obtained by integrating both sides of the modified equation (4) during a very short time interval.

3.1.4. A first, very simple example

Let us illustrate on a very basic example, the grounding ideas of the algebraic approach. For this purpose consider the first order linear system:

[4] It is reminiscent to that the most practitioners in electronics are doing.

[5] Let us stress again that many computer simulations and several laboratory experiments have been already successfully achieved and can be quite favorably compared with the existing techniques.

[6] Especially in signal processing.

$$\dot{y}(t) = ay(t) + u(t) + \gamma_0, \tag{60}$$

where a is an unknown parameter to be identified and γ_0 is an unknown constant perturbation. With the notations of operational calculus and $y_0 = y(0)$, equation (5) reads:

$$s\widehat{y}(s) = a\widehat{y}(s) + \widehat{u}(s) + y_0 + \frac{\gamma_0}{s} \tag{61}$$

where $\widehat{y}(s)$ represents the Laplace transform of $y(t)$.

In order to eliminate the term γ_0, multiply first the both hand-sides of this equation by s and next take their derivatives with respect to s:

$$\frac{d}{ds}\left[s\left\{s\widehat{y}(s) = a\widehat{y}(s) + \widehat{u}(s) + y_0 + \frac{\gamma_0}{s}\right\}\right] \tag{62}$$

$$\Rightarrow 2s\widehat{y}(s) + s^2\widehat{y}'(s) = a\left(s\widehat{y}'(s) + \widehat{y}(s)\right) + s\widehat{u}'(s) + \widehat{u}(s) + y_0. \tag{63}$$

Recall that $\widehat{y}'(s) \triangleq \frac{d\widehat{y}(s)}{ds}$ corresponds to $-ty(t)$. Assume $y_0 = 0$ for simplicity of presentation [7]. Then for any $\nu > 0$,

$$s^{-\nu}\left[2s\widehat{y}(s) + s^2\widehat{y}'(s)\right] = s^{-\nu}\left[a(s\widehat{y}'(s) + \widehat{y}(s)) + s\widehat{u}'(s) + \widehat{u}(s)\right]. \tag{64}$$

For $\nu = 3$, we obtained the estimated value a:

$$a = \frac{2\int_0^T d\lambda \int_0^\lambda y(t)dt - \int_0^T ty(t)dt + \int_0^T d\lambda \int_0^\lambda tu(t)dt - \int_0^T d\lambda \int_0^\lambda d\sigma \int_0^\sigma u(t)dt}{\int_0^T d\lambda \int_0^\lambda d\sigma \int_0^\sigma y(t)dt - \int_0^T d\lambda \int_0^\lambda ty(t)dt} \tag{65}$$

Since $T > 0$ can be very small, estimation *via* (10) is very fast.

Note that equation (10) represents an on-line algorithm, which involves only two kinds of operations on u and y: (1) multiplications by t, and (2) integrations over a pre-selected time interval.

If we now consider an additional noise of zero mean in (5), say:

$$\dot{y}(t) = ay(t) + u(t) + \gamma_0 + n(t), \tag{66}$$

it can be considered as a fast fluctuating signal. The order ν in (9) determines the order of iterations in the integrals (3 integrals in (10)). Those iterated integrals are low-pass filters which are attenuating the fluctuations.

This example, even simple, clearly demonstrates how algebraic techniques proceed:

- they are algebraic: operations on s-functions;
- they are non-asymptotic: parameter a is obtained from (10) in a finite time;
- they are deterministic: no knowledge of the statistical properties of the noise n is required.

[7] If $y_0 \neq 0$ one has to take above derivatives of order 2 with respect to s, in order to eliminate the initial condition.

3.1.5. A second simple example, with delay

Consider the first order, linear system with constant input delay [8]:

$$\dot{y}(t) + ay(t) = y(0)\delta + \gamma_0 H + bu(t-\tau). \tag{67}$$

Here we use a distributional-like notation, where δ denotes the Dirac impulse and H is its integral, i.e. the Heaviside function (unit step) [9]. Still for simplicity, we suppose that the parameter a is known. The parameter to be identified is now the delay τ. As previously, γ_0 is a constant perturbation, a, b, and τ are constant parameters. Consider also a step input $u = u_0 H$. A first order derivation yields:

$$\ddot{y} + a\dot{y} = \varphi_0 + \gamma_0\,\delta + b\,u_0\,\delta_\tau, \tag{68}$$

where δ_τ denotes the delayed Dirac impulse and $\varphi_0 = (\dot{y}(0) + ay(0))\,\delta + y(0)\,\delta^{(1)}$, of order 1 and support $\{0\}$, contains the contributions of the initial conditions. According to Schwartz theorem, multiplication by a function α such that $\alpha(0) = \alpha'(0) = 0$, $\alpha(\tau) = 0$ yields interesting simplifications. For instance, choosing $\alpha(t) = t^3 - \tau\,t^2$ leads to the following equalities (to be understood in the distributional framework):

$$\begin{array}{rcl} t^3\,[\ddot{y} + a\dot{y}] & = & \tau\,t^2\,[\ddot{y} + a\dot{y}], \\ bu_0\,t^3\delta_\tau & = & bu_0\,\tau\,t^2\delta_\tau. \end{array} \tag{69}$$

The delay τ becomes available from $k \geq 1$ successive integrations (represented by the operator H), as follows:

$$\tau = \frac{H^k(w_0 + a\,w_3)}{H^k(w_1 + a\,w_2)}, \qquad t > \tau, \tag{70}$$

where the w_i are defined using the notation $z_i = t^i\,y$ by:

$$\begin{array}{rcl} w_0 & = & t^3\,y^{(2)} = -6\,z_1 + 6\,z_2^{(1)} - z_3^{(2)}, \\ w_1 & = & t^2\,y^{(2)} = -2\,z_0 + 4\,z_1^{(1)} - z_2^{(2)}, \\ w_2 & = & t^2\,y^{(1)} = 2\,z_1 - z_2^{(1)}, \\ w_3 & = & t^3\,y^{(1)} = 3\,z_2 - z_3^{(1)}. \end{array}$$

These coefficients show that $k \geq 2$ integrations avoid any derivation in the delay identification.

Figure 1 gives a numerical simulation with $k = 2$ integrations and $a = 2, b = 1, \tau = 0.6$, $y(0) = 0.3, \gamma_0 = 2, u_0 = 1$. Due to the non identifiability over $(0, \tau)$, the delay τ is set to zero until the numerator or the denominator in the right hand side of (15) reaches a significant nonzero value.

Again, note the realization algorithm (15) involves two kinds of operators: (1) integrations and (2) multiplications by t. It relies on the measurement of y and on the knowledge of a. If a is also unknown, the same approach can be utilized for a simultaneous identification of a and τ. The following relation is derived from (14):

[8] This example is taken from [93]. For further details, we suggest the reader to refer to it.

[9] In this document, for the sake of simplicity, we make an abuse of the language since we merge in a single notation the Heaviside function H and the integration operator. To be rigorous, the iterated integration (k times) corresponds, in the operational domain, to a division by s^k, whereas the convolution with H (k times) corresponds to a division by $s^k/(k-1)!$. For $k = 0$, there is no difference and $H * y$ realizes the integration of y. More generally, since we will always apply these operations to complete equations (left-and right-hand sides), the factor $(k-1)!$ makes no difference.

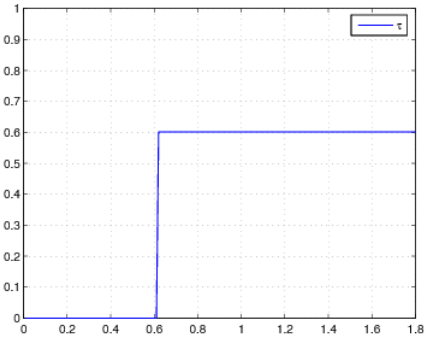


Figure 1. Delay τ identification from algorithm (15)

$$\tau(H^k w_1) + a\,\tau(H^k w_2) - a\,(H^k w_3) = H^k w_0, \tag{71}$$

and a linear system with unknown parameters $(\tau, a\,\tau, a)$ is obtained by using different integration orders:

$$\begin{pmatrix} H^2 w_1 & H^2 w_2 & H^2 w_3 \\ H^3 w_1 & H^3 w_2 & H^3 w_3 \\ H^4 w_1 & H^4 w_2 & H^4 w_3 \end{pmatrix} \begin{pmatrix} \widehat{\tau} \\ \widehat{a\tau} \\ -\widehat{a} \end{pmatrix} = \begin{pmatrix} H^2 w_0 \\ H^3 w_0 \\ H^4 w_0 \end{pmatrix}.$$

The resulting numerical simulations are shown in Figure 2 . For identifiability reasons, the obtained linear system may be not consistent for $t < \tau$.

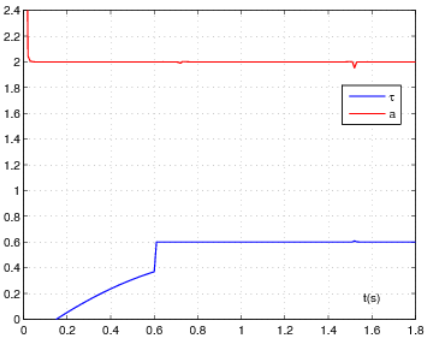


Figure 2. Simultaneous identification of a and τ from algorithm (16)

3.2. Finite time estimation of derivatives

Numerical differentiation, i.e. determining the time derivatives of various orders of a noisy time signal, is an important but difficult ill-posed theoretical problem. This fundamental issue has attracted a lot of attention in many fields of engineering and applied mathematics (see, e.g. in the recent control literature [94], [95], [109], [108], [115], [116], and the references therein).

3.2.1. Model-free techniques for numerical differentiation

A common way of estimating the derivatives of a signal is to resort to a least squares fitting and then take the derivatives of the resulting function. In [119], [117] this problem was revised through our algebraic approach. The approach can be briefly explained as follows:

- The coefficients of a polynomial time function are linearly identifiable. Their estimation can therefore be achieved as above. Indeed, consider a real-valued polynomial function $x_N(t) = \sum_{\nu=0}^{N} x^{(\nu)}(0)\frac{t^\nu}{\nu!} \in \mathbb{R}[t]$, $t \geq 0$, of degree N. Rewrite it in the well-known notations of operational calculus:

$$X_N(s) = \sum_{\nu=0}^{N} \frac{x^{(\nu)}(0)}{s^{\nu+1}}.$$

  Here we use $\frac{d}{ds}$, which corresponds in the time domain to the multiplication by $-t$. Multiply both sides by $\frac{d^\alpha}{ds^\alpha} s^{N+1}$, $\alpha = 0, 1, \cdots, N$. The quantities $x^{(\nu)}(0)$, $\nu = 0, 1, \cdots, N$ are given by the triangular system of linear equations:

$$\frac{d^\alpha s^{N+1} X_N}{ds^\alpha} = \frac{d^\alpha}{ds^\alpha}\left(\sum_{\nu=0}^{N} x^{(\nu)}(0) s^{N-\nu}\right). \tag{72}$$

  The time derivatives, i.e. $s^\mu \frac{d^\iota X_N}{ds^\iota}$, $\mu = 1, \cdots, N$, $0 \leq \iota \leq N$ are removed by multiplying both sides of Equation (17) by $s^{-\overline{N}}$, $\overline{N} > N$.

- For an arbitrary analytic time function, let us apply the preceding calculations to a suitable truncated Taylor expansion. Consider a real-valued analytic time function defined by the convergent power series $x(t) = \sum_{\nu=0}^{\infty} x^{(\nu)}(0)\frac{t^\nu}{\nu!}$, where $0 \leq t < \rho$. Approximate $x(t)$ in the interval $(0, \varepsilon)$, $0 < \varepsilon \leq \rho$ by its truncated Taylor expansion $x_N(t) = \sum_{\nu=0}^{N} x^{(\nu)}(0)\frac{t^\nu}{\nu!}$ of order N. Introduce the operational analogue of $x(t)$, i.e. $X(s) = \sum_{\nu \geq 0} \frac{x^{(\nu)}(0)}{s^{\nu+1}}$. Denote by $[x^{(\nu)}(0)]_{e_N}(t)$, $0 \leq \nu \leq N$, the numerical estimate of $x^{(\nu)}(0)$, which is obtained by replacing $X_N(s)$ by $X(s)$ in Eq. (17). It can be shown [104] that a good estimate is obtained in this way.

Thus using elementary differential algebraic operations, we derive an explicit formulae yielding pointwise derivative estimation for each given order. Interesting enough, it turns out that the Jacobi orthogonal polynomials [129] are inherently connected with the developed algebraic numerical differentiators. A least-squares interpretation then naturally follows [118], [119] and this leads to a key result: the algebraic numerical differentiation is as efficient as an appropriately chosen time delay. Though, such a delay may not be tolerable in some real-time applications. Moreover, instability generally occurs when introducing delayed signals in a control loop. Note however that since the delay is known *a priori*, it is always possible to derive a control law, which compensates for its effects (see [127]). A second key feature of the algebraic numerical differentiators is its very low complexity which allows for a real-time implementation. Indeed, the n^{th} order derivative estimate (that can be directly managed for $n \geq 2$, without using n cascaded estimators) is expressed as the output of the linear time-invariant filter, with finite support impulse response $h_{\kappa,\mu,n,r}(\cdot)$. Implementing such a stable and causal filter is easy and simple. This is achieved either in continuous-time or in discrete-time when only discrete-time samples of the observation are available. In the latter case, we obtain a tapped delay line digital filter by considering any numerical integration method with equally-spaced abscissas.

3.2.2. Model-based estimation of derivatives

If we assume that the derivatives to be estimated are unmeasured states of a process that generates the signal, then the differentiation techniques can be considered as left invertibility algorithms. In this sense, the previous algebraic estimation achieves a "model-free" left inversion. Now, when such a model is available, the *finite-time observers* relying on higher order sliding modes [123] and homogeneity properties [124], [120] also represent possible non-asymptotic algorithms for differentiation [10]. Using such model-based techniques appears to be complementary[11]. The left-inversion results have been already obtained for several classes of models: linear systems [106], nonlinear systems [92], delay systems [2] and hybrid systems [114].

[10] Usually, observer design yields asymptotic convergence of the estimation error dynamics. The main advantages of such a technique in the case of linear systems are simplicity of design, estimation with a filtering action and global stability property. Nevertheless, the filtering property is not ensured for nonlinear systems and the stability property is generally obtained only locally. For these reasons, in the case of nonlinear systems, finite-time observers and estimators have been proposed in the literature [116], [124], [125], [105]...

[11] The choice between the two approaches will be done after comparison with respect to the indicators 1–3, and taking into account the application (for instance, the system bandwidth, system dimension), the kind of discontinuity, the observer in the control loop or not...

OPALE Project-Team

3. Scientific Foundations

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 since its very beginning. A software integration platform has been designed by the OPALE project 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 [70].

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. Scientific Foundations

3.1. Mathematical analysis and simulation of wave propagation

Our activity relies on the existence of mathematical models established by physicists to model the propagation of waves in various situations. The basic ingredient is a partial differential equation (or a system of partial differential equations) of the hyperbolic type that are often (but not always) linear for most of the applications we are interested in. The prototype equation is the wave equation:

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = 0,$$

which can be directly applied to acoustic waves but which also constitutes a simplified scalar model for other types of waves (This is why the development of new numerical methods often begins by their application to the wave equation). Of course, taking into account more realistic physics will enrich and complexify the basic models (presence of sources, boundary conditions, coupling of models, integro-differential or non linear terms,...)

It is classical to distinguish between two types of problems associated with these models: the time domain problems and the frequency domain (or time harmonic) problems. In the first case, the time is one of the variables of which the unknown solution depends and one has to face an evolution problem. In the second case (which rigorously makes sense only for linear problems), the dependence with respect to time is imposed a priori (via the source term for instance): the solution is supposed to be harmonic in time, proportional to $e^{i\omega t}$, where $\omega > 0$ denotes the pulsation (also commonly, but improperly, called the frequency). Therefore, the time dependence occurs only through this pulsation which is given a priori and plays the rôle of a parameter: the unknown is only a function of space variables. For instance, the wave equation leads to the Helmholtz wave equation (also called the reduced wave equation) :

$$-c^2 \Delta u - \omega^2 u = 0.$$

These two types of problems, although deduced from the same physical modelling, have very different mathematical properties and require the development of adapted numerical methods.

However, there is generally one common feature between the two problems: the existence of a dimension characteristic of the physical phenomenon: the wavelength. Intuitively, this dimension is the length along which the searched solution varies substantially. In the case of the propagation of a wave in an heterogeneous medium, it is necessary to speak of several wavelengths (the wavelength can vary from one medium to another). This quantity has a fundamental influence on the behavior of the solution and its knowledge will have a great influence on the choice of a numerical method.

Nowadays, the numerical techniques for solving the basic academic and industrial problems are well mastered. A lot of companies have at their disposal computational codes whose limits (in particular in terms of accuracy or robustness) are well known. 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. A large part of research in mathematics 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 treatment of more and more complex problems (non local models, non linear models, coupled systems, ...)

- the study of specific phenomena or features such as guided waves, resonances,...
- the development of approximate models in various situations,
- imaging techniques and inverse problems related to wave propagation.

REALOPT Project-Team

3. Scientific Foundations

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 Ilog-CPLEX or FICO/Dash-Optimization's Xpress-mp). 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 the study of non-linear models (non linearities are inherent in some engineering, economic and scientific applications) where solution techniques build on the best MIP approaches while demanding much more than simple extensions. Robust optimization is another area 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 mathematical programming (polyhedral approaches, Dantzig-Wolfe decomposition, non-linear integer programing, 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* [59]. 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, *cutting plane procedures* defined in the previous section reveal to be powerful. 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. Constraint Programming (CP)

Constraint Programming 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 (minimizing the maximum of several variable values). Mixed Integer Programming (MIP), on the other hand, is 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 example, some scheduling and timetabling problems are tightly constrained and have a sum-type objective. For such problems, it is reasonable to use algorithms that exploit complementary strengths of Constraint Programming and Mixed Integer Programming.

3.5. Mixed Integer NonLinear Programming (MINLP)

Many engineering, management, and scientific applications involve not only discrete decisions, but also nonlinear relationships that significantly affect the feasibility and optimality of solutions. MINLP problems

combine the difficulties of MIP with the challenges of handling nonlinear functions. MINLP is one of the most flexible modeling paradigms available. However, solving such models is much more challenging: available softwares are not nearly as effective as standard softwares for linear MIP. The most powerful algorithms combine sophisticated methods that maintain outer linear programming approximation or convex relaxations with branch-and-bound enumeration; hence, the role of strong convex reformulations is crucial. The development of results for structured sub-models are essential building blocks. Preprocessing and bound reduction (domain reduction logic similar to that used in CP) are quite important too. Finally, decomposition methods also permit to develop tight outer approximations.

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. Scientific Foundations

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 [30], 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:

$$\alpha_X(t_0,\omega) = \sup\left\{\alpha : \limsup_{\rho\to 0} \sup_{t,u\in B(t_0,\rho)} \frac{|X_t - X_u|}{\rho^\alpha} < \infty\right\}, \tag{73}$$

and

$$\widetilde{\alpha}_X(t_0,\omega) = \sup\left\{\alpha : \limsup_{\rho\to 0} \sup_{t,u\in B(t_0,\rho)} \frac{|X_t - X_u|}{\|t-u\|^\alpha} < \infty\right\}. \tag{74}$$

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 [5]). 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 [16]. 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 [50]. 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 [23], [31]. See [64], [66] 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 [16], 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 [72].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 [27], [51], [60]. 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, [40]) 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 $q \in \mathbb{R}$,

$$S_n(q) = \sum_{k=0}^{n-1} |Y_n^k|^q, \tag{75}$$

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 [61], [63]. Let us also mention the multifractality of TCP traffic, that has been both observed experimentally and proved on simplified models of TCP [2], [47].

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 [28] 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 [49] 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 [42]. 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 [7]. Other natural phenomena displaying a relation between amplitude and exponent include temperatures records and RR intervals extracted from ECG [38].

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:

$$\overline{\underline{X}}_{\alpha'}^{\beta'} = \alpha' + (\beta'-\alpha')\frac{X-\min_K(X)}{\max_K(X)-\min_K(X)}$$

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:

$$\begin{aligned}\Lambda_{\alpha',\beta'} : \mathcal{C}([0,1],[\alpha,\beta]) &\to \mathcal{C}([0,1],[\alpha,\beta])\\ Z(.) &\mapsto \overline{\underline{B(.,g(Z(.))}}_{\alpha'}^{\beta'}\end{aligned}$$

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 [74]. 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.

More formally, a multistable process is a process which is, at each time u, tangent to a stable process [59]. Recall that a process Y is said to be tangent at u to the process Y'_u if:

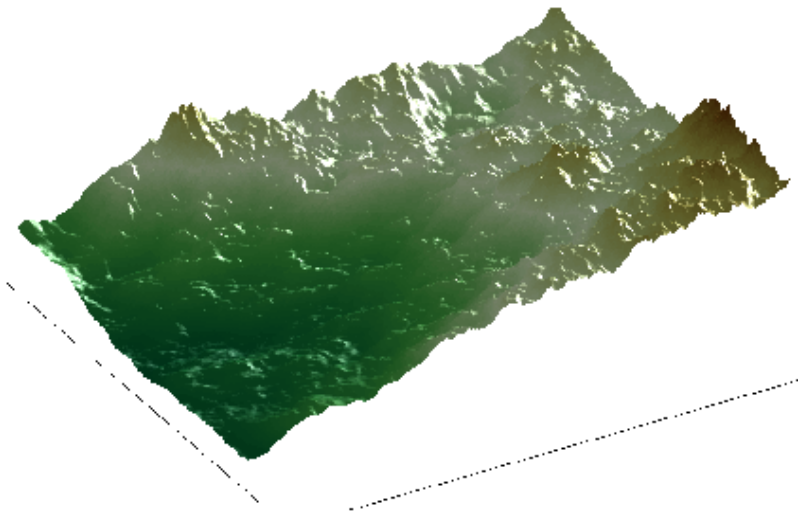


Figure 1. Self-regulating miltifractional process with $g(x) = 1 - x^2$

$$\lim_{r \to 0} \frac{Y(u + rt) - Y(u)}{r^h} = Y'_u(t), \tag{76}$$

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 [28]: 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 [9], [8]. 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:

$$Y(t) = a(t) \sum_{(\mathsf{X},\mathsf{Y}) \in \Pi} \mathsf{Y}^{<-1/\alpha(t)>} \left(|t - \mathsf{X}|^{h(t)-1/\alpha(t)} - |\mathsf{X}|^{h(t)-1/\alpha(t)} \right) \quad (t \in \mathbb{R}). \tag{77}$$

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 [75]. 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.

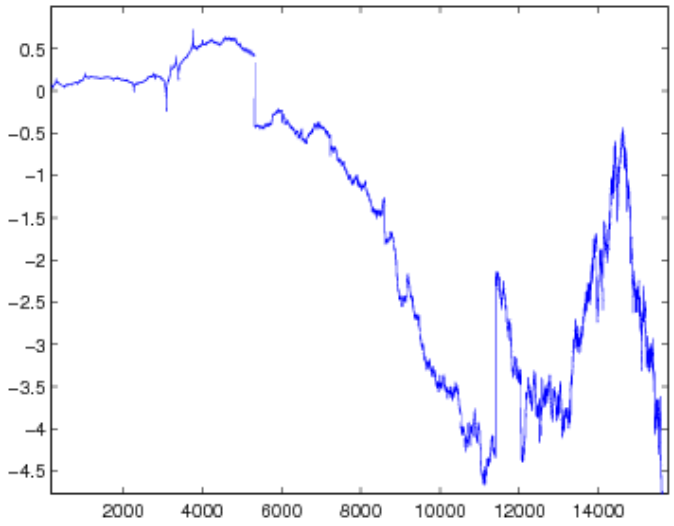


Figure 2. Linear multistable multifractional motion with linearly increasing α and linearly decreasing H

Multiparameter processes

In order to use stochastic processes to represent the variability of multidimensional phenomena, it is necessary to define extensions for indices in $\mathbb{R}^N$ ($N \geq 2$) (see [67] for an introduction to the theory of multiparameter processes). Two different kinds of extensions of multifractional Brownian motion have already been considered: an isotropic extension using the Euclidean norm of $\mathbb{R}^N$ and a tensor product of one-dimensional processes on each axis. We refer to [13] for a comprehensive survey.

These works have highlighted the difficulty of giving satisfactory definitions for increment stationarity, Hölder continuity and covariance structure which are not closely dependent on the structure of $\mathbb{R}^N$. For example, the Euclidean structure can be unadapted to represent natural phenomena.

A promising improvement in the definition of multiparameter extensions is the concept of *set-indexed processes*. A set-indexed process is a process whose indices are no longer "times" or "locations" but may be some compact connected subsets of a metric measure space. In the simplest case, this framework is a generalization of the classical multiparameter processes [62]: usual multiparameter processes are set-indexed processes where the indexing subsets are simply the rectangles $[0, t]$, with $t \in \mathbb{R}^N_+$.

Set-indexed processes allow for greater flexibility, and should in particular be useful for the modeling of censored data. This situation occurs frequently in biology and medicine, since, for instance, data may not be constantly monitored. Censored data also appear in natural terrain modeling when data are acquired from sensors in presence of hidden areas. In these contexts, set-indexed models should constitute a relevant frame.

A set-indexed extension of fBm is the first step toward the modeling of irregular phenomena within this more general frame. In [18], the so-called *set-indexed fractional Brownian motion (sifBm)* was defined as the mean-zero Gaussian process $\{\mathbf{B}^H_U;\ U \in \mathcal{A}\}$ such that

$$\forall U, V \in \mathcal{A}; \quad E[\mathbf{B}_U^H \ \mathbf{B}_V^H] = \frac{1}{2}\left[m(U)^{2H} + m(V)^{2H} - m(U \bigwedge V)^{2H}\right]$$

where $\mathcal{A}$ is a collection of connected compact subsets of a measure metric space and $0 < H \leq \frac{1}{2}$.

This process appears to be the only set-indexed process whose projection on increasing paths is a one-parameter fractional Brownian motion [17]. The construction also provides a way to define fBm's extensions on non-euclidean spaces, *e.g.* indices can belong to the unit hyper-sphere of $\mathbb{R}^N$. The study of fractal properties needs specific definitions for increment stationarity and self-similarity of set-indexed processes [20]. We have proved that the sifBm is the only Gaussian set-indexed process satisfying these two (extended) properties.

In the specific case of the indexing collection $\mathcal{A} = \{[0,t], t \in \mathbb{R}_+^N\} \cup \{\varnothing\}$, the sifBm can be seen as a multiparameter extension of fBm which is called *multiparameter fractional Brownian motion (MpfBm)*. This process differs from the Lévy fractional Brownian motion and the fractional Brownian sheet, which are also multiparameter extensions of fBm (but do not derive from set-indexed processes). The local behaviour of the sample paths of the MpfBm has been studied in [12]. The self-similarity index H is proved to be the almost sure value of the local Hölder exponent at any point, and the Hausdorff dimension of the graph is determined in function of H.

The increment stationarity property for set-indexed processes, previously defined in the study of the sifBm, allows to consider set-indexed processes whose increments are independent and stationary. This generalizes the definition of Bass-Pyke and Adler-Feigin for Lévy processes indexed by subsets of $\mathbb{R}^N$, to a more general indexing collection. We have obtained a Lévy-Khintchine representation for these set-indexed Lévy processes and we also characterized this class of Markov processes.

SCIPORT Team

3. Scientific Foundations

3.1. Automatic Differentiation

Participants: Laurent Hascoët, Valérie Pascual.

automatic differentiation (AD) Automatic transformation of a program, that returns a new program that computes some derivatives of the given 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.

checkpointing General trade-off technique, used in adjoint-mode AD, that trades duplicate execution of a part of the program to save some memory space that was used to save intermediate results. Checkpointing a code fragment amounts to running this fragment without any storage of intermediate values, thus saving memory space. Later, when such an intermediate value is required, the fragment is run a second time to obtain the required values.

Automatic or Algorithmic Differentiation (AD) differentiates *programs*. An AD tool takes as input a source program P that, given a vector argument $X \in I\!R^n$, computes some vector result $Y = F(X) \in I\!R^m$. The AD tool generates a new source program P' that, given the argument X, computes some derivatives of F. 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

$$\begin{array}{lcl} P & \text{is} & \{I_1; I_2; \cdots I_p;\}, \\ F & = & f_p \circ f_{p-1} \circ \cdots \circ f_1, \end{array} \tag{78}$$

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 chain rule gives the Jacobian of F

$$F'(X) = f'_p(X_{p-1}) \,.\, f'_{p-1}(X_{p-2}) \,.\, \cdots \,.\, f'_1(X_0) \tag{79}$$

which can be mechanically written as a sequence of instructions I'_k. Combining the I'_k with the control of P yields P'. This can be generalized to higher level derivatives, Taylor series, etc.

In practice, the Jacobian $F'(X)$ is often too expensive to compute and store, but most applications only need projections of $F'(X)$ such as:

- **Sensitivities**, defined for a given direction $\dot{X}$ in the input space as:

$$F'(X).\dot{X} = f'_p(X_{p-1}) \,.\, f'_{p-1}(X_{p-2}) \,.\, \cdots \,.\, f'_1(X_0) \,.\, \dot{X} \quad . \tag{80}$$

  Sensitivities are easily computed from right to left, interleaved with the original program instructions. This is the *tangent mode* of AD.

- **Adjoints**, defined for a given weighting $\overline{Y}$ of the outputs as:

$$F'^*(X).\overline{Y} = f_1'^*(X_0).f_2'^*(X_1).\ \cdots\ .f_{p-1}'^*(X_{p-2}).f_p'^*(X_{p-1}).\overline{Y} \quad . \tag{81}$$

Adjoints are most efficiently computed from right to left, because matrix×vector products are cheaper than matrix×matrix products. This is the *adjoint mode* of AD, most effective for optimization, data assimilation [33], adjoint problems [28], or inverse problems.

Adjoint-mode AD turns out to make a very efficient program, at least theoretically [30]. 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 regularly extend our models and tools to handle these new constructs.

3.2. Static Analysis and Transformation of programs

Participants: Laurent Hascoët, Valérie Pascual.

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, contain each a list of instructions to be executed in sequence, and whose arcs 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 simultaneously, 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 whether a variable is initialized or not, at any location in the source program.

data dependence analysis Program analysis that studies the itinerary of values during program execution, from the place where a value is generated to the places where it is used, and finally to the place where it is overwritten. The collection of all these itineraries is often stored 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 conversely 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 [20] to improve the quality of the produced code. 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.
- 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 [23], [31], [21]. 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 are 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 [24], 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* [25] 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. In addition to the very theoretical limit of *undecidability*, there are practical limitations to how much information one can infer from programs that use arrays [37], [26] or pointers. In general, conservative *over-approximations* are always made that lead to derivative code that is less efficient than possibly achievable.

3.3. Automatic 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 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 [28], 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 [27] 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 [32], [35] to reduce memory usage.

If for instance the model is steady, one tradeoff can use the iterated states in the direct order [29]. Another tradeoff can use only the fully converged final state. Since tradeoff approaches can be error-prone, we advocate incorporating them into the AD model and into the AD tools.

SELECT Project-Team

3. Scientific Foundations

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. Scientific Foundations

3.1. Introduction

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 [70].

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 $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to I\!R$ 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 [1] $r_1, r_2, ...$ 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.

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:

[1] 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.

$$V^{\pi}(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = x, \pi\right], \tag{82}$$

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 [66]) 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) [64], 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 [65]. 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 $x \in \mathcal{X}$,

$$V^*(x) = \max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{83}$$

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:

$$\pi^*(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{84}$$

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 ([76]):

- 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 challenge 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 [71], 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.* [63] 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 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 depend 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 [68]. 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 [67] 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:

$$F(x) = \int_{\Theta} f(x|\theta) G(d\theta) = \sum_{k=1}^{\infty} \omega_k f(x|U_k) \tag{85}$$

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:

$$\begin{aligned} f_{k+1|k}(X|Z^{(k)}) &= \int f_{k+1|k}(X|W) f_{k|k}(W|Z^{(k)}) \delta W \\ f_{k+1|k+1}(X|Z^{(k+1)}) &= \frac{f_{k+1}(Z_{k+1}|X) f_{k+1|k}(X|Z^{(k)})}{\int f_{k+1}(Z_{k+1}|W) f_{k+1|k}(W|Z^{(k)}) \delta W} \end{aligned} \tag{86}$$

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 $\mathcal{X}$; [2]
- $Z_{k+1} = \{z_1, ..., z_m\}$ 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.

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.

[2] The state x_i of a target is usually composed of its position, its velocity, etc.

SIERRA Project-Team

3. Scientific Foundations

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.

SIMPAF Project-Team

3. Scientific Foundations

3.1. General framework

Partial Differential Equations, Kinetic Equations, Conservation Laws, Hyperbolic Systems, Fluid Mechanics, Parabolic Systems, Computational Fluid Dynamics, Plasma Physics, Asymptotic analysis

The scientific activity of the project is concerned with Partial Differential Equations (PDE) arising from the physical description of particles and fluids. It covers various viewpoints:

- At first, the words "particles and fluids" could simply mean that we are interested independently in models for particles, which can either be considered as individuals (which leads to "N-particle models", N ranging from 1 to many) or through a statistical description (which leads to kinetic equations) as well as in models for fluids like Euler and Navier-Stokes equations or plasma physics.
- However, many particle systems can also be viewed as a fluid, via a passage from microscopic to macroscopic viewpoint, that is, a hydrodynamic limit.
- Conversely, a fruitful idea to build numerical solvers for hyperbolic conservation laws consists in coming back to a kinetic formulation. This approach has motivated the introduction of the so-called kinetic schemes.

By nature these problems describe multiscale phenomena and one of the major difficulties when studying them lies in the interactions between the various scales: number of particles, size, different time and length scales, coupling...
The originality of the project is to consider a wide spectrum of potential applications. In particular, the word "particles" covers various and very different physical situations and it has evolved with the composition of the team. One may think of:

- charged particles: description of semi-conductor devices or plasmas;
- bacteria, individuals or genes as in models motivated by biology or population dynamics;
- droplets and bubbles, as in Fluid/Particles Interaction models which arise in the description of sprays and aerosols, smoke and dust, combustion phenomena (aeronautics or engine design), industrial process in metallurgy...
- cross-links in polymer chains to describe rubber elasticity;
- oxyde molecules to model corrosion phenomena at the miroscopic scale and derive effective macroscopic equations;
- cold atoms...

We aim at focusing on all the aspects of the problem:

- Modelling mathematically complex physics requires a deep discussion of the leading phenomena and the role of the physical parameters. With this respect, the asymptotic analysis is a crucial issue, the goal being to derive reduced models which can be solved with a reduced numerical cost but still provide accurate results in the physical situations that are considered.
- The mathematical analysis of the equations provides important qualitative properties of the solutions: well-posedness, stability, smoothness of the solutions, large time behavior... which in turn can motivate the design of numerical methods.
- Eventually, we aim at developing specific numerical methods and performing numerical simulations for these models, in order to validate the theoretical results and shed some light on the physics.

The team has been composed in order to study these various aspects simultaneously. In particular, we wish to keep a balance between modelling, analysis, development of original codes and simulations.

3.2. Interactions of Micro- and Macroscopic Scales and Simulations

Statistical Physics, Homogenization, Asymptotic Preserving Schemes

3.2.1. Homogenization methods

Homogenization methods aim at replacing a PDE with highly oscillatory coefficients by an effective PDE with smoother coefficients, whose solution captures the averaged behavior of the true oscillatory solution. The effective determination of the homogenized PDE is however not trivial (especially in the nonlinear or/and stochastic cases). Numerical approximations of the solution of the homogenized PDE is the heart of numerical homogenization.

Homogenization methods are used in many application fields. The two applications we are specifically interested in are material sciences (in particular the determination of macroscopic constitutive laws for rubber starting from polymer-chain networks) and nuclear waste storage (in particular the evolution of nuclear wastes in complex storage devices).

The team in interested in qualitative as well as quantitative results, and theoretical as well as numerical results. Challenging questions are mainly related to nonlinear problems (nonlinear elasticity for instance) and stochastic problems (especially regarding quantitative results).

3.2.2. Statistical physics : molecular dynamics

The team is concerned with the numerical simulation of stochastically perturbed Molecular Dynamics. The main goal is to handle in the same simulation the fastest time scales (e.g. the oscillations of molecular bindings), and the slowest time scales (e.g. the so-called reaction coordinates). Recently, M. Rousset co-authored a monograph [64] which summarizes standard and state-of-the-art free energy calculations, that are used to accelerate slow variables in MD simulations.

3.2.3. Statistical physics: dynamical friction, fluctuations and approach to equilibrium

In models of charge transport, say transport of electrons, a phenomenological friction force is generally introduced, which is proportional to the velocity v. The dissipation induced by such a term is essential for the description of phenomena such as Ohm's law and approach to equilibrium. Our idea is to go back to a microscopic framework, with a description of the energy exchanges between the electrons and the surrounding medium which is the ultimate source of the dissipation of energy by the medium and of an effective friction force. We have shown numerically and argued theoretically that the balance between the fluctuations and the dissipation by the medium drives the particle to thermal equilibrium. The goal is now to provide rigorous proof of this statement. As a first step in this program, results will be obtained in an appropriate weak coupling limit. This program requires efforts in modelling, probability and analysis, but the questions are also really challenging for numerics, due, notably, to the large number of degrees of freedom involved in the equation. The subject is at the heart of the PhD work of É. Soret, now in her second year as a PhD student.

3.2.4. Cold Atoms

In the framework of the Labex CEMPI, C. Besse, S. De Bièvre and G. Dujardin are working, in collaboration with J.-C. Garreau and the cold-atom team at PhLAM, on the mathematical analysis and the numerical simulation of kicked rotor systems. Such systems are experimentally realized at PhLAM. A triple goal is being pursued: understand the effect of non-linearities on dynamical localization, understand dynamical localization in systems other than kicked rotors, and exploring the limits of the analogy between kicked systems and the Anderson model.

3.3. Plasma

In the context of the Galileo satellite-positioning system, C. Besse and C. Yang, members of the ANR Iodissee project, developed a hierarchy of plasma models which describe ionospheric scintillations. This hierarchy involves many small parameters, and they introduced an asymptotic preserving scheme which allows one to take small parameters into account without solving the problem on a very fine grid [8]. The next step is to

understand the fading and phase variations when waves propagate in this medium. This is a work in progress with P. Lafitte and S. Minjeaud (CNRS, Université de Nice).

3.4. Finite element and finite volume methods

Conservation Laws, Anti-Diffusive Schemes, Viscous Flows, Control, Turbulence, Finite element methods, Finite volume methods

3.4.1. Control in Fluid Mechanics

Flow control techniques are widely used to improve the performances of planes or vehicles, or to drive some internal flows arising for example in combustion chambers. Indeed, they can sensibly reduce energy consumption, noise disturbances, or prevent the flow from undesirable behaviors.
E. Creusé is involved in the development of open and closed active flow control, with applications to recirculation in engines or blood flows.

3.4.2. Numerical Methods for Viscous Flows

Numerical investigations are very useful to check the behavior of systems of equations modelling very complicate dynamics. In order to simulate the motion of mixtures of immiscible fluids having different densities, a recent contribution of the team was to develop an hybrid Finite Element / Finite Volume scheme for the resolution of the variable density 2D incompressible Navier-Stokes equations. The main points of this work were to ensure the consistency of the new method [56] as well as its stability for high density ratios [54]. In order to answer these questions, we have developed a MATLAB code and a C++ code. In the following of this work, C. Calgaro and E. Creusé now have in mind the following objectives, in collaboration with T. Goudon (team COFFEE, Inria Sophia-Antipolis) :

- Distribute the matlab version of the code (with an accurate documentation and a graphic interface) to promote new collaborations in the domain and compare alternative numerical solution methods (for instance to compare updating LU factorizations, see [55]);
- To generalize the stability results obtained in [54] for the scalar transport equation to the full 2D Euler system, in particular very low density values density (near vacuum);
- Complete the C++ code to treat more general hydrodynamic models (combustion theory, transport of pollutants). We plan to check the behavior of the equations (typically the Kazhikhov-Smagulov model of powder-snow avanlanches) in the regime when the current existence theory does not apply, and extend our kinetic asymptotic-based schemes to such problems.

3.4.3. A posteriori error estimators for finite element methods

A posteriori estimates, finite element methods

The team works on a posteriori error estimators for finite element methods, applied to the resolution of several partial differential equations. The objective is to derive useful tools in order to control the global error between the exact solution and the approximated one (reliability of the estimator), and to control the local error leading to adaptive mesh refinement strategies (efficiency of the estimator).
More specifically, E. Creusé works on the derivation of some "reconstruction estimators" based on gradient averaging, for diffusion problems (with S. Nicaise, LAMAV, Valenciennes), the Reissner-Mindlin system (PhD of É. Verhille), and the Maxwell equations (PhD of Z. Tang).

3.5. Numerical analysis of Schrödinger equations

Dispersive equations, Schrödinger equations

3.5.1. Modelling of quantum dot-helium

In collaboration with G. Reinish (Nice Observatory) and V. Guðmundsson (University of Reykjavik), C. Besse and G. Dujardin are working on the numerical computation of the ground state and the first bound states of the non linear Schrödinger-Poisson system with confining quadratic potential in 2 space dimensions. This models quatum dot helium (*i.e.* the behavior of a pair of quantum electrons in a strong confining potential). The goal is to perform after that numerical time stepping methods to simulate the dynamics of the NLSP system and compute accurately some quantities of physical interest as functions of time, in order to be able to compare the competition between the Coulomb (repulsive) interaction and the binding (attractive) forces due to the confinement in this model as well as in other quantum mechanics models.

3.5.2. Dispersive Schrödinger-like equations

In collaboration with M. Taki (PhLAM laboratory, Lille), C. Besse and G. Dujardin are considering dispersive equations modelling the propagation of a laser beam in an optical fiber. They are trying to explain the possible ways of creating rogue waves in the propagation of laser beams. More generally, they are trying to explain which terms in the dispersive Schrödinger-like equations obtained by the physicists allow which physical behaviour of the solutions (e.g. the creation of rogue waves).

TAO Project-Team

3. Scientific Foundations

3.1. Introduction

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 or 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 others evolve and adapt to new topics.

The first one, OPT-SIG, addresses stochastic continuous optimization, taking 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 best continuous stochastic optimizer, CMA-ES, ranging from convergence analysis (Youhei Akimoto's and Verena Heidrich-Meister's post-docs) to a rigorous benchmarking methodology. Incidentally, this benchmark platform has been acknowledged since 2008 as "the" international continuous optimization benchmark, and its extension is at the core of the ANR project NumBBO (starting end 2012). Another priority is to address the current limitations of CMA-ES in terms of high-dimensional or expensive optimization (respectively Ouassim Ait El Hara's and Ilya Loshchilov's PhDs). Mouadh Yagoubi and Zyed Bouzarkouna (resp. CIFRE PSA and IFP-EN) PhD's have continued the EC tradition of industrial breakthrough applications, on which the EC fame solidly relies.
The second SIG, UCT-SIG, benefits from the MoGo expertise 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 one-year stay in Taiwan, and witnessed by the FP7 STREP Citines, the pending ADEME Post contract, and the Ilab with Artelys.
A third SIG, DIS-SIG, is devoted to the modeling and optimization of (large scale) distributed systems. DIS-SIG pursues and extends the goals of the former *Autonomic Computing* SIG, initiated by Cécile Germain 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. More generally, how to model and manage network-based activities has been acknowledged a key topic *per se* in the last months, including: i) the modeling of multi-agent systems and the exploitation of simulation results in the SimTools RNSC network frame; ii) the management of the core communication topology for distributed SAT solving, in the *Microsoft-TAO project* framework. Further extensions are planned 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.

A fourth 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 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. Indeed hyper-parameter tuning is still present in TAO, chiefly for continuous optimization (OPT-SIG, section 3.3) and AI planning (CRI-SIG, section 3.4).

3.2. Optimal Decision Making under Uncertainty

Participants: Olivier Teytaud [correspondent], Jean-Joseph Christophe, Adrien Couëtoux, Hassen Doghmen, Jérémie Decock, Nicolas Galichet, Manuel Loth, Marc Schoenauer, Michèle Sebag.

The UCT-SIG works on sequential optimization problems, where a decision has to be made at each time step along a finite time horizon, and the underlying problem involves uncertainties along an either adversarial or stochastic setting.

Application domains include energy management at various time scales and more generally planning, on the one hand, and games (Go, MineSweeper, NoGo [12]) on the other hand.

The main advances done this year include:

- In the domain of computer Go some new performances have been realized [16] and survey papers have been published in Communications of the ACM [10] and as a chapter [65].
- The extension of Upper Confidence Trees to continuous or large domains (states and/or actions) and to domains with high expertise or strong structure has been done [37], [31], [38].
- The extension of Upper Confidence Trees to multi-objective settings has been done, improving on scalarization-based multi-objective reinforcement learning [56].
- Anytime algorithms for discrete time control (note that the classical stochastic dynamic programming is by no means anytime) have been developed and integrated in the Metis software (section 5.1), based on the above cited results.
- Hybrid approaches combining upper confidence trees and e.g. direct policy search or domain specific approaches to yield robust performance w.r.t. long-term effects and take advantage of the combinatorial structure of the domain have been designed, specifically including problem-specific expertise in the playout phase in the domain of job-shop scheduling [53] or MineSweeper [54].
- Optimization algorithms for direct policy search have been designed [55].
- Within the European STREP Mash project, our favorite tools (in particular Monte-Carlo Tree Search) have been extended to difficult settings with no possibility to "undo" a decision [63]. The notion of "risk of exploration" has been investigated [59].
- An experimental analysis of bandit algorithms for small budget cases [36] got the excellent paper award at TAAI 2012.
- In collaboration with Christian Shulte (KTH, Stockholm), one of the main contributors to the well-known general-purpose CP solver *GECODE* (http://www.gecode.org/), and within the Microsoft-Inria joint lab- Adapt project, ideas from UCT have been integrated in GECODE and applied to the job-shop scheduling problem with good first results [52].
- The optimization of low-discrepancy sequences has been done, improving on the best results so far [7]; note that low-discrepancy sequences have been exploited in quite a few of our past works.

3.3. Continuous Optimization

Participants: Ouassim Ait ElHara, Yohei Akimoto, Anne Auger, Zyed Bouzarkouna, Alexandre Chotard, Nikolaus Hansen, Ilya Loshchilov, Verena Heidrich-Meisner, Yann Ollivier, Marc Schoenauer, Michèle Sebag, Olivier Teytaud, Mouadh Yagoubi.

Our main expertise in continuous optimization is on stochastic search algorithms. We address theory, algorithm design and applications. The methods we investigate are adaptive techniques able to learn iteratively parameters of the distribution used to sample solutions. The Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is nowadays one of the most powerful methods for derivative-free continuous optimization. We work on different variants of the CMA-ES to improve it in various contexts as described below. We have proven the convergence of simplified variants of the CMA-ES algorithm using the theory of stochastic approximation providing the first proofs of convergence on composite of twice continuously differentiable functions and used Markov chain analysis for analyzing the step-size adaptation rule of the CMA-ES algorithm.

New algorithms based on surrogates and for constrained optimization. A new variant of CMA-ES to address constraint optimization has been designed [22]. In the Zyed Bouzarkouna's PhD, defended in April 2012, a CIFRE PhD in cooperation with IFP-EN (Institut Français du Pétrole - Energies Nouvelles), new variants of CMA-ES coupled with local-meta-models for expensive optimization have been proposed. They have been applied to well placement problem in oil industry [2]. In the context of his PhD thesis (to be defended in January 2013), Ilya Loshchilov has proposed different surrogate variants of CMA-ES based on ranking-SVM that preserve the invariance to monotonic transformation of the CMA-ES algorithm [51]. He has also explored new restart mechanisms for CMA-ES [47].

Benchmarking. We have continued our effort for improving standards in benchmarking and pursued the development of the COCO (COmparing Continuous Optimizers) platform. We have organized the ACM GECCO 2012 workshop on Black-Box-Optimization Benchmarking [2] and benchmarked different variants of the CMA-ES algorithms [27] [30] [29] [28] [48] [49] [50]. Our new starting ANR project NumBBO, centered on the COCO platform, aims at extending it for large-scale, expensive, constrained and multi-objective optimization.

Theoretical proofs of convergence. We have defined and analyzed a variant of CMA-ES being able to prove that its covariance matrix converges to the inverse Hessian on convex-quadratic functions [18]. We have analyzed the convergence of continuous time trajectories associated to step-size adaptive Evolution Strategies on monotonic C^2-composite Functions and proved the local convergence towards local minima [19]. This work has been the starting point for analyzing convergence of step-size adaptive ESs using the framework of stochastic approximation (work about to be submitted). In the context of his thesis, Alexandre Chotard has analyzed the step-size adaptation algorithm of CMA-ES on linear function using the theory of Markov Chains [35].

Multi-objective optimization. Mouadh Yagoubi has completed his PhD [3], a CIFRE cooperation with PSA (Peugeot-Citroen automotive industry). His work addressed the multi-disciplinary multi-objective optimization of a diesel engine, that led to propose some asynchronous parallelization of expensive objective functions (more than 2 days per evaluation for the complete 3D model) [57].

3.4. Designing criteria

Participants: Jamal Atif, Nicolas Bredèche, Cyril Furtlehner, Yoann Isaac, Victorin Martin, Jean-Marc Montanier, Hélène Paugam-Moisy, Marc Schoenauer, Michèle Sebag.

[2] see http://coco.gforge.inria.fr/doku.php?id=bbob-2012

This new SIG, rooted on the claim that *What matters is the criterion*, aims at defining new learning or optimization objectives reflecting fundamental properties of the model, the problem or the expert prior knowledge.

A statistical physics perspective. In the context of the ANR project TRAVESTI (http://travesti.gforge.inria.fr) which ended this year, we have worked to address more specifically the inverse pairwise Markov random field (MRF) model. On one side we have formalized how the Ising model [64] can be used to perform travel time inference in this context (submitted to AISTATS). Extending the ordinary linear response theory in the vicinity of the so called "Bethe reference point" instead of the interaction-free reference point [67], we were able to provide explicit and tractable formulas to the Plefka's expansion and the related natural gradient at that point. These can be used into various possible algorithms for generating approximate solutions to the inverse Ising problem which we do investigate now. In parallel we have proposed also in [67], a method based on the "iterative proportional scaling" to learn both the factor graph and the couplings by selecting links one by one (submitted to AISTATS). In the Gaussian MRF case this can be implemented efficiently due to local transformations of the precision matrix after adding one link, which is unfortunately not possible in the Ising MRF. It is competitive with L_0 based approach in terms of precision and computational cost, while incidentally the L_1 based method potentially cheaper is not working well for this problem. The flexibility of the method offers in addition the possibility to combine it with spectral constraints like walk-summability with belief propagation or/and graph structure constrain to enforce compatibility with BP. With no additional computational cost we get a complete set of good trade-offs between likelihood and compatibility with BP. Concerning the analysis of the belief propagation we establish in [60] some sufficient condition for encoding a set of local marginals into a stable belief propagation fixed point. Following some work of last year concerning the modeling of congestion at the microscopic level we have finalized in [9] the analysis of a new family of queuing processes where the service rate is coupled stochastically to the number of clients leading a large deviation formulation of the fundamental diagram of traffic flow.

Multi-objective AI Planning. Within the ANR project DESCARWIN (http://descarwin.lri.fr), Mostepha-Redouane Kouadjia worked on the **multi-objective approach** to AI Planning using the Evolutionary Planner *Divide-and-Evolve*, that evolves a sequential decomposition of the problem at hand: each sub-problem is then solved in turn by some embedded classical planner [72]. Even though the embedded planner is single-objective, DaE can nevertheless handle multi-objective problems: Current work includes the implementation of the multi-objective version of DaE, the definition of some benchmark suite, and some first numerical experiments, comparing in particular the results of a full Pareto approach to those of the classical aggregation method. These works resulted in 3 conference papers recently accepted, introducing a tunable benchmark test suite [45], demonstrating that the best quality measure for parameter tuning in this multi-objective framework is the hypervolume, even in the case of the aggregation approach [46], and comparing the evolutionary multi-objective approach with the aggregation method, the only method known to the AI Planning community [44]. The parameter-tuning algorithm designed for DAE, called Learn-and-Optimize, was published in the selection of papers from *Evolution Artificielle* conference [26]. Though originally designed for Evolutionary AI Planning, the method is applicable to domains where instances sharing similar characteristics w.r.t parameter tuning can be grouped in domains.

Image understanding. Sequential image understanding refers to the decision making paradigm where objects in an image are successively segmented/recognized following a predefined strategy. Such an approach generally raises some issues about the "best" segmentation sequence to follow and/or how to avoid error propagation. Within the new sequential recognition framework proposed in [8], these issues are addressed as the objects to segment/recognize are represented by a model describing the spatial relations between objects. The process is guided by a criterion derived from visual attention, specifically a saliency map, used to optimize the segmentation sequence. Spatial knowledge is also used to ensure the consistency of the results and to allow backtracking on the segmentation order if needed. The proposed approach was applied for the segmentation of internal brain structures

in magnetic resonance images. The results show the relevance of the optimization criteria and the relevance of the backtracking procedure to guarantee good and consistent results. In [70] we propose a method for simultaneously segmenting and recognizing objects in images, based on a structural representation of the scene and on a constraint propagation method. Within the ANR project LOGIMA, our goal is to address sequential object recognition as an abduction process [69]. Similar principles are at the core of Yoann Isaac's PhD (Digiteo Unsupervised Brain project), in collaboration with CEA LIST. The dictionary-learning approach used to decompose the EEG signal is required to comply with the structure of the data (e.g. spatio-temporal continuity; submitted).

Robotic value systems. Within the European SYMBRION IP, a key milestone toward autonomous cognitive agents has been to provide robots with internal or external rewards, yielding an interesting or competent behavior. Firstly, an objective-free setting referred to as open-ended evolution has been investigated [17], [5], where the criterion to be optimized is left implicit in the reproduction process. in the Secondly, preference-based reinforcement learning has been investigated in Riad Akrour's PhD, where the robot demonstrations are assessed by the expert and these assessments are used to learn a model of the expert's expectations. In [21], this work has been extended and combined with active learning to yield state-of-the art performances with few binary feedbacks from the expert. The hormone-based neural net controller first proposed by T. Schmickl et al. has been thoroughly analyzed and simplified in collaboration with Artificial Life Laboratory from Graz [34].

3.5. Distributed systems

Participants: Cécile Germain-Renaud [correspondent], Philippe Caillou, Dawei Feng, Nadjib Lazaar, Michèle Sebag.

The DIS-SIG explores the issues related to modeling and optimizing distributed systems, ranging from very large scale computational grids to multi-agent systems and distributed constraint solvers.

Coping with non-stationarity. Most existing work on modeling the dynamics of grid behavior assumes a steady-state system and concludes to some form of long-range dependence (slowly decaying correlation) in the associated time-series. But the physical (economic and sociological) processes governing the grid behavior dispel the stationarity hypothesis. When the behavior can be modeled as a time series, an appealing class of models is a sequence of stationary processes separated by break points.The optimisation problem for structural break detection is difficult, because of high dimensionality and a complex objective function. Then, evolutionary algorithms are a method of choice. [15] revisits the optimisation strategy in the light of the general advances in evolutionary computation and the specific opportunities for a separable representation. The single-level optimisation problem is decoupled into a bilevel optimisation. The upper level is the problem of finding the optimal number and location of the break points. The lower level optimises the autoregressive models given the number and locations of break points. At the upper level, our optimisation strategy exploits the state-of-the-art CMA-ES (Covariance Matrix Adaptation - Evolutionary Strategy) instead of the relatively straightforward Genetic Algorithm proposed in the classic AutoPARM fitting procedure for non-stationary time series. The associated representation addresses an important shortcoming of : the distance in the chromosomes space better maps to the distance in the model space. Furthermore, the representation becomes scalable. More precisely, it scales linearly with the length of the data set, independently of the cost of the objective function.

Fault management. Isolating users from the inevitable faults in large distributed systems is critical to Quality of Experience. Thus a significant part of the software infrastructure of large scale distributed systems collects information that will be exploited to discover if, where, and when the system is faulty. In the context of end-to-end probing as the class of monitoring techniques, minimizing the number of probes for a given discovery performance target is critical. While detection and diagnosis have the obvious advantage of providing an explanation of the failure, by exhibiting culprits, they strongly rely on a priori knowledge that is not available for massively distributed

systems. Thus [39], [62] formulates the problem of probe selection for fault prediction based on end-to-end probing as a Collaborative Prediction (CP) problem, based on the reasonable assumption of an underlying factorial model. On an extensive experimental dataset from the EGI grid, the combination of the Maximum Margin Matrix Factorization approach to CP and Active Learning shows excellent performance, reducing the number of probes typically by 80% to 90%.

Multi-agent and games. The main research focus concerning multi-agent systems was on the observation and automatic description of multi-agent based simulations. Whereas usual parameter space exploration systems observe several experiments, the increasing complexity of simulations makes it harder to understand and describe what happens during a single experiment. It is simple to define global indicators to have an overview of the simulation or to follow individual agents, but the most interesting phenomena often occur at an intermediate level, where *groups of agents* are found. The group level is also suited to analyse and display the dynamics of the model. The online and agent-oriented analysis was achieved and its statistical soundness was assessed [6]. In collaboration with the CEA LIST laboratory, we developed a generic tool (SimAnalyer), which can be used online (with NetLogo) of offline (with Logs) to identify, describe, follow [33] and reproduce [58] clusters of agents in a simulation. To select the most interesting clusters and descriptive variables, new activity indicators reflecting the simulation dynamics have been designed [13].

Parallel SAT Solving. Recent Parallel SAT solvers use the so-called Conflict-Directed Clause Learning to exchange clauses between the different cores. However, when the number of cores increases, systematic clause sharing leads to communication saturation. Nadjib Lazaar's post-doc, funded by the Microsoft-Inria joint lab, investigated how the communication topology can be optimized online using a Multi-Armed Bandit setting [68], with an improvement of circa 10% (in number of problems solved) and over 50% (in computational time) over ManySAT 2.0, on the 2012 SAT and UNSAT problem suite.

TOSCA Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

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. Scientific Foundations

3.1. Introduction

The research conducted by ABS focuses on two main directions in Computational Structural Biology (CSB), each such direction calling for specific algorithmic developments. These directions are:
- Modeling interfaces and contacts,
- Modeling the flexibility of macro-molecules.

3.2. Modeling Interfaces and Contacts

Docking, interfaces, protein complexes, structural alphabets, scoring functions, Voronoi diagrams, arrangements of balls

Problems addressed. 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 [1], 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 [46]. Yet, in spite of almost three decades of investigations, the basic principles guiding the formation of interfaces and accounting for its stability are unknown [49]. Current investigations follow two routes. From the experimental perspective [31], 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 [43]. From the modeling perspective, the main issue consists of guessing the hot residues from sequence and/or structural informations [38].

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 [2], 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 [23], 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 [47], [34]. 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 [39]. In doing so, the energy defined is nothing but the Kullback-Leibler divergence between the distributions $\{p_i\}$ and $\{q_i\}$.

Methodological developments. Describing interfaces poses problems in two settings: static and dynamic.

[1] 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.

[2] 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 [10]. 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 [24]. At the atomic level and in spite of recent observations on the local structure of the neighborhood of a given atom [48], 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 [30].

A structural alphabet at the atomic level may be seen as an alphabet featuring for an atom of a given type all the conformations this atom may engage into, depending on its neighbors. One way to tackle this problem consists of extending the notions of molecular surfaces used so far, so as to encode multi-body relations between an atom and its neighbors [8]. In order to derive such alphabets, the following two strategies are obvious. On 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.

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 [42]. 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

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 [22]. 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 [21], 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 [20], 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 [20]. 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.

3.3.3. Methodological developments

As outlined by the previous discussion, a number of methodological developments are called for. On the experimental side, the problem of fostering the interpretation of data is under scrutiny. Of particular interest is the disambiguation of proteomics signals (TAP data, mass spectrometry data), and that of density maps coming from electron microscopy. 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. The second one encompasses assessment tools, in order to single out the reconstructions which best comply with the experimental data.

3.4. Modeling the Flexibility of Macro-molecules

Folding, docking, energy landscapes, induced fit, molecular dynamics, conformers, conformer ensembles, point clouds, reconstruction, shape learning, Morse theory

Problems addressed. 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 and Molecular Dynamics simulations generate ensembles of conformations, called `conformers`. 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 [3]. 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.

Methodological developments. At the side-chain level, the question of improving rotamer libraries is still of interest [29]. 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) [45]. 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 [41], to Morse theory [36] and to analysis of meta-stable states of time series [37] have been proposed.

[3] Assuming local forces are prominent, which in turn subsumes electrostatic interactions are not prominent.

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 [40].

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 [6]. 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 [25]; the study of Morse-like constructions stemming from distance functions to points [33]; the analysis of topological invariants of the model and the samples, and their comparison [26], [27].

Last but not least, gaining insight on such questions would also help to effectively select a reduced set of conformations best representing a larger number of conformations. This selection problem is indeed faced by flexible docking algorithms that need to maintain and/or update collections of conformers for the second stage of the *diffusion - conformer selection - induced fit* complex formation model.

AMIB Project-Team

3. Scientific Foundations

3.1. RNA and protein structures

3.1.1. RNA

Participants: Julie Bernauer, Alain Denise, Rasmus Fonseca, Feng Lou, Yann Ponty, Mireille Régnier, Philippe Rinaudo, Jean-Marc Steyaert.

Common activity with P. Clote (Boston College and Digiteo).

3.1.1.1. From RNA structure to function

We are currently developing a combinatorial approach, based on random generation, to design small and structured RNAs. An application of such a methodology to the Gag-Pol HIV-1 frameshifting site will be carried out with our collaborators at IGM. We hope that, upon capturing the hybridization energy at the design stage, one will be able to gain control over the rate of frameshift and consequently fine-tune the expression of *Gag/Pol*. Our goal is to build these RNA sequences such that their hybridization with existing mRNAs will be favorable to independent folding, and will therefore affect the stability of some secondary structures involved in recoding events. Moreover it has been observed, mainly on bacteria, that some mRNA sequences may adopt alternate folds. Such events are called a conformational switch, or riboswitch. A common feature of recoding events and riboswitches is that some structural element on the mRNA initiates and unusual action of the ribosome, or allows for an alternate fold under some environmental conditions. One challenge is to predict genes that might be subject to riboswitches.

3.1.1.2. Beyond secondary structure

One of our major challenges is to go beyond secondary structure. Over the past decade, few attempts have been made to predict the 3D structure of RNA from sequence only. So far, few groups have taken this leap. Despite the promises shown by their preliminary results, these approaches currently suffer to a limiting scale due to either their high algorithmic complexity or their difficult automation. Using our expertise in algorithmics and modeling, we plan to design original methods, notably within the AMIS-ARN project (ANR BLANC 2008-2012) in collaboration with PRISM at Versailles University and E.Westhof's group at Strasbourg.

1. *Ab initio* modeling: Starting from the predicted RNA secondary structure, we aim to detect *local structural motifs* in it, giving local 3D conformations. We use the resulting partial structure as a flexible scaffold for a multi-scale reconstruction, notably using game theory. We believe the latter paradigm offers a more realistic view of biological processes than global optimization, used by our competitors, and constitutes a real originality of our project.
2. Comparative modeling: we investigate new algorithms for predicting 3D structures by a comparative approach. This involves comparing multiple RNA sequences and structures at a large scale, that is not possible with current algorithms. Successful methods must rely both on new graph algorithms and on biological expertise on sequence-structure relations in RNA molecules.

3.1.1.3. RNA 3D structure evaluation

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. Their function is mainly determined by the structure those molecules adopt as protein and nucleic acids differ from polypeptides and polynucleotides by their spatial organization. This is specially challenging for RNA where structure flexibility is key.

To address those issues, one has to explore the biologically possible spatial configurations of a macromolecule. The two most common techniques currently used in computational structural biology are Molecular Dynamics (MD) and Monte Carlo techniques (MC). Those techniques require the evaluation of a potential or force-field, which for computational biology are often empirical. They mainly consist of a summation of bonded forces associated with chemical bonds, bond angles, and bond dihedrals, and non-bonded forces associated with van der Waals forces and electrostatic charges. Even if there exists implicit solvent models, they are yet not very well performing and still require a lot of computation time.

Our goal, in collaboration with the Levitt lab at Stanford University and H. van den Bedem at the Stanford Synchrotron Radiation Laboratory (Associate Team ITSNAPhttp://pages.saclay.inria.fr/julie.bernauer/EA_ITSNAP/) is to develop knowledge-based (KB) potentials, based on measurements on known RNA 3D structures and provide sampling for experimental structure fitting and docking conformation generation. KB potential are quick to evaluate during a simulation and can be used without having to explicitly address the solvent problem. They can be developed at various levels of representation: -atom, base, nucleotide, domain- and could allow the modelling of a wide size range: from a hairpin to the whole ribosome. We also intend to combine these knowledge-based potentials with other potentials (hybrid modelling) and template-based techniques, allowing accurate modelling and dynamics study of very large RNA molecules. Such studies are still a challenge. We will also study conformations for experimental data fitting by extension the innovative, robotics-inspired Kino-Geometric Sampler conformational search algorithm for proteins to nucleic acids and to include experimental data. KGS models a protein as a kinematic linkage and additionally considers hydrogen bonds that "close" kinematic cycles. In closed kinematic cycles rotatable bonds can no longer be deformed independently without breaking closure. KGS preserves all kinematic cycles, and thus hydrogen bonds, by sampling in a subspace of conformational space defined by all closure constraints. KGS exhibits a singularly large search radius and optimally reduces the number of free parameters. These unique features enable flexible 'docking' of atomic models in the data while moderating the risk of overfitting at low resolution. The KGS procedure can also accommodate knowledge-based potentials to improve evaluation of putative conformations and their interactions (see below).

3.1.2. PROTEINS

Participants: Jérôme Azé, Julie Bernauer, Adrien Guilhot-Gaudeffroy, Jean-Marc Steyaert.

3.1.2.1. Docking and evolutionary algorithms

As mentioned above, the function of many proteins depends on their interaction with one or many partners. Docking is the study of how molecules interact. Despite the improvements due to structural genomics initiatives, the experimental solving of complex structures remains a difficult problem. The prediction of complexes, *docking*, proceeds in two steps: a configuration generation phase or *exploration* and an evaluation phase or *scoring*. As the verification of a predicted conformation is time consuming and very expensive, it is a real challenge to reduce the time dedicated to the analysis of complexes by the biologists. Various algorithms and techniques have been used to perform exploration and scoring [49]. The recent rounds of the CAPRI challenge show that real progress has been made using new techniques [46], [3]. Our group has strong experience in cutting edge geometric modelling and scoring techniques using machine learning strategies for protein-protein complexes. In a collaboration with A. Poupon, INRA-Tours, a method that sorts the various potential conformations by decreasing probability of being real complexes has been developed. It relies on a ranking function that is learnt by an evolutionary algorithm. The learning data are given by a geometric modelling of each conformation obtained by the docking algorithm proposed by the biologists. Objective tests are needed for such predictive approaches. The *Critical Assessment of Predicted Interaction*, CAPRI, a community wide experiment modelled after CASP was set up in 2001 to achieve this goal (http://www.ebi.ac.uk/msd-srv/capri/). First results achieved for CAPRI'02 suggested that it is possible to find good conformations by using geometric information for complexes. This approach has been followed (see section New results). As this new algorithm will produce a huge amount of conformations, an adaptation of the ranking function learning step is needed to handle them. In the near future, we intend to extend our approach to protein-RNA complexes.

Such as in the protein case, the function of RNA molecules also depends on their interaction with one or many partners. Upon interaction, RNA molecules often undergo large conformation changes. Understanding how these molecules interact with proteins would allow better targeting for therapeutic studies. The CAPRI (Critical Assessment of PRediction of Interactions) challenge1 has shown that classical docking procedures largely fail when large conformation change occurs and when RNA is involved. This is especially true for RNA molecules, whose large-scale dynamics remain often unknown. Modeling RNA conformational changes is made hard by the inherent flexible nature of their structure but also by the electrostatics involved. These are hard to model and often lead to computationally expensive simulations. Even if for small RNA molecules, molecular dynamics can be used, such simulations are hard to extend to larger molecules and protein-RNA complexes.

For many diseases, such as cancer and HIV, microRNA molecules play a very important role regulating gene expression by guiding the RISC. Some miRNAs have been shown to suppress tumors and are thus ideal candidates for the development of therapeutic agents. Even if various computational techniques have been developed to predict miRNA targets, none of these consider the structural aspects of the interactions between components of the RISC and miRNA. We aim to target these problems, in collaboration with the Huang lab at HKUST (PHC Procore)

The combination of Voronoi models at a coarse-grained level and powerful machine learning techniques allows the accurate scoring of protein-protein complexes [12]. Our actual machine learning approach for proteins is a combination of several machine learning approaches (evolutionary algorithm, decision trees, decision rules,...). By adapting these approaches to protein-RNA, we would have a fast and efficient technique for scoring large protein-RNA complexes where conformational changes are involved.

Working with RNA instead of protein introduce many major differences in the machine learning approaches. RNA conformations are often smaller than protein conformations, which has an impact on the values of the descriptors used to describe objects. Due to the size differences between RNA and protein, it is often more difficult to generate (during the modeling stage) conformations closed to the biological solution (near native solution). The machine learning algorithms therefore need to take into account all theses specificities to be able to learn good predictive models from data that are not very close to the real solution.

The acquired knowledge on RNA flexibility, dynamics and the importance of the sequence will be a strong advance in the modeling of protein-RNA interactions we are working on. Il will help the development of scoring functions based on Voronoi models for RNA and provide us with the level of flexibility needed in complex conformational search. We also intend to develop hybrid KB potentials for complexes from hybrid RNA KB data. These could be incorporated in leading-edge flexible docking modeling software such as Rosetta.

3.2. Annotations

3.2.1. Word counting

Participants: Alain Denise, Daria Iakovishina, Yann Ponty, Mireille Régnier, Jean-Marc Steyaert.

We aim at enumerating or generating sequences or structures that are *admissible* in the sense that they are likely to possess some given biological property. Team members have a common expertise in enumeration and random generation of combinatorial structures. They have developped computational tools for probability distributions on combinatorial objects, using in particular generating functions and analytic combinatorics. Admissibility criteria can be mainly statistic; they can also rely on the optimisation of some biological parameter, such as an energy function.

The ability to distinguish a significant event from statistical noise is a crucial need in bioinformatics. In a first step, one defines a suitable probabilistic model (null model) that takes into account the relevant biological properties on the structures of interest. A second step is to develop accurate criteria for assessing (or not) their exceptionality. An event observed in biological sequences, is considered as exceptional, and therefore biologically significant, if the probability that it occurs is very small in the null model. Our approach to

compute such a probability consists in an enumeration of good structures or combinatorial objects. Thirdly, it is necessary to design and implement efficient algorithms to compute these formulae or to generate random data sets. Typical examples that motivate research on words and motifs counting are *Transcription Factor Binding Sites*, TFBSs, consensus models of recoding events and some RNA structural motifs. When relevant motifs do not resort to regular languages, one may still take advantage of combinatorial properties to define functions whose study are amenable to our algebraic tools. One may cite secondary structures and recoding events.

Fast development of high throughput technologies has generated a new challenge for computational biology. The main bottlenecks in applications are the computational analysis of experimental data.

As a first example, numerous new assembling algorithms have recently appeared. Still, the comparison of the results arising from these different algorithms led to significant differences for a given genome assembly. Clearly, strong constraints from the underlying technologies, leading to different data (size, confidence,...) are one origin of the problems and a deeper interpretation is needed, in order to improve algorithms and confidence in the results. One objective is to develop a model of errors, including a statistical model, that takes into account the quality of data for the different technologies, and their volume. This is the subject of an international collaboration with V. Makeev's lab (IoGene, Moscow) and MAGNOME project-team. Second, Next Generation Sequencing open the way to the study of structuralvariants in the genome, as recently described in [44]. Defining a probabilistic model that takes into account main dependencies -such as the GC content- is a task o D. Iakovishina's thesis, in a starting 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. Stochastic context-free grammars (SCFG's) have long been used to model both structural and statistical properties of genomic sequences, particularly for predicting the structure of sequences or for searching for motifs. They can also be used to generate random sequences. However, they do not allow the user to fix the length of these sequences. We developed algorithms for random structures generation that respect a given probability distribution on their components. Our approach is based on the concept of *weighted* combinatorial classes, in combination with the so-named *recursive* method for generating combinatorial structures. To that purpose, one first translates the (biological) structures into combinatorial classes, using the *symbolic method*, an algebraic framework developed by Flajolet *et al*. Adding weights to the atoms allows one to bias the probabilities towards the desired distribution. The main issue is to develop efficient algorithms for finding the suitable weights. An implementation was given in the `GenRGenS` software http://www.lri.fr/~genrgens/, and a generic optimizer that automatically derives suitable parameters for a given grammar, is currently being developped.

In 2005, a new paradigm appeared in the *ab initio* secondary structure prediction [45]: 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 [48], [4], 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 [47].

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 [10], [21].

3.2.3. Programmed -1 ribosomal frameshifting

Participants: Patrick Amar, Jérôme Azé, Alain Denise, Christine Froidevaux, Yann Ponty, Cong Zeng.

During protein synthesis, the ribosome decodes the mRNA by assigning a specific amino acid to each codon or nucleotide triplet. Throughout this process the ribosome moves along the mRNA molecule three nucleotides at a time. However, encounters of specific signals found in mRNA from many viruses lead the ribosome to shift one nucleotide backward thus changing its reading frame. We aim at developing a new computational approach that is able to detect these specific signals in genomic databases in order to better understand the molecular choreography leading to the ribosomal frameshifting, which ultimately will help to rationally design new antiviral drugs. As candidates sequences are expected to be numerous, we aim at developing a ranking method to identify the most relevant sequences. Biological testing of these most promising identified candidates by our collaborators from IGM will help us to refine our computational method.

3.2.4. Knowledge extraction

Participants: Jérôme Azé, Jiuqiang Chen, Sarah Cohen-Boulakia, Christine Froidevaux.

Our main goal is to design semi-automatic methods for annotation. A possible approach is to focus on the way we could discover relevant motifs in order to make more precise links between function and motifs sequence. For instance, a commonly accepted hypothesis is that function depends on the order of the motifs present in a genomic sequence. Likewise we must be able to evaluate the quality of the annotation obtained. This necessitates giving an estimate of the reliability of the results. This may use combinatorial tools described above. It includes a rigorous statement of the validity domain of algorithms and knowledge of the results provenance. We are interested in provenance resulting from workflow management systems that are important in scientific applications for managing large-scale experiments and can be useful to calculate functional annotations. A given workflow may be executed many times, generating huge amounts of information about data produced and consumed. Given the growing availability of this information, there is an increasing interest in mining it to understand the difference in results produced by different executions.

3.2.5. Systems Biology

Participants: Patrick Amar, Sarah Cohen-Boulakia, Alain Denise, Christine Froidevaux, Loic Paulevé, Sabine Peres, Mireille Régnier, Jean-Marc Steyaert.

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).

3.2.5.1. Simulations and behavior analysis for metabolism modeling

A great number of methods have been proposed for the study of the behavior of large biological systems. Two methods have been developed and are in use in the team, depending on the specific problems under study : the first one is based on a discrete and direct simulation of the various interactions between the reactants, while the second one deals with an abstract representation by means of differential equations from which we extract various types of features pertaining to the system.

We investigate on the computational modelling step of the cycle by developing a computer simulation system, HSIM, that mimics the interactions of biomolecules in an environment modelling the membranes and compartments found in real cells. In collaboration with biologists from the AMMIS lab. at Rouen we have used HSIM to show the properties of grouping the enzymes of the phosphotransferase system and the glycolytic pathway into metabolons in *E. coli*. In another collaboration with the SYSDIAG Lab (UMR 3145) at Montpellier, we participate at the CompuBioTic project. This is a Synthetic Biology project in the field of medical diagnosis: its goal is to design a small vesicle containing specific proteins 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 human pathologies. We used HSIM to help the design and to test qualitatively and quantitatively this *"biological computer"* before *in vitro*.

Given the set of biochemical reactions which 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. We have developed a software `Mpas` (Metabolic Pathway Analyser Software) that renders the translation in terms of a systems of o.d.e's automatic; then the simulations are also made easy and almost automatic. 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 particular we have been able to prove mathematically the decisive influence of the enzyme PEMT on the Choline/Ethylamine cycles (M. Behzadi's thesis, defended in 2011).

3.2.5.2. Comparison of Metabolic Networks

In the context of a national project, 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 will be studied to find new enzymes allowing the transformation of biomass. The KEGG database contains pathways related to fungi and other species. By analysing these known pathways with rules mining approaches, we would be able to predict new enzymes activities.

3.2.5.3. Signalling networks

AMIB and INRA-BIOS (A. Poupon, Tours) are partners in a two years project ASAM (2011-2012). This project aims to help the understanding of signalling pathways involving G protein-coupled receptors (GPCR) which 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. The aim of ASAM is thus to provide means to semi-automatically construct signalling networks of GPCRS.

ASCLEPIOS Project-Team

3. Scientific Foundations

3.1. Introduction

Tremendous progress has been made in the automated analysis of biomedical images during the past two decades [91]. Readers who are neophyte to the field of medical imaging will find an interesting presentation of acquisition techniques of the main medical imaging modalities in [82], [80]. Regarding the target applications, a good review of the state of the art can be found in the book *Computer Integrated Surgery* [78], in N. Ayache's article [86] and in the more recent syntheses [87] [91]. The scientific journals *Medical Image Analysis* [73], *Transactions on Medical Imaging* [79], and *Computer Assisted Surgery* [81] 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) [76], [77] or ISBI'2010 (Int. Symp. on Biomedical Imaging) [75].

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 [92], [106]. 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 [109], or to measure some functional properties of the heart from dynamic sequences of Magnetic Resonance [85], Ultrasound or Nuclear Medicine images [93].

Despite these advances and successes, one can notice that statistical models of the anatomy are still very crude, resulting in poor registration results in deformable regions of the body, or between different subjects. If some algorithms exploit the 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 the observed images and signals, but also more efficient tools to detect anomalies, predict evolutions, simulate and assess 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 the images are multidimensional (3 spatial coordinates and possibly one temporal dimension), but medical protocols tend to include multi-sequence (or multi-parametric) [1] and multi-modal images [2] for each single patient.

[1] 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 acquisition (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 to measure for instance the direction of white matter fibers in the brain (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 subtle higher T2* signal which can be detected with sophisticated image processing techniques.

[2] Multimodal acquisition consists in acquiring on the same patient images from 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 on the physics of image acquisition and observed tissues, as well as on 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 3.4 and 3.5 .

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. Biological Image Analysis

In biology, a huge number of images of living systems are produced every day to study the basic mechanisms of life and pathologies. If some bio-imaging *principles* are the same as the ones used for medical applications (e.g. MR, CT, US, PET or SPECT), the bio-imaging *devices* are usually customized to produce images of higher resolution [3] for the observation of small animals (typically rodents). In addition, Optical Imaging (OI) techniques and biophotonics are developing very fast. This includes traditional or Confocal Microscopy (CM), multi-photon confocal microscopy, Optical Coherent Tomography (OCT), near-infrared imaging, diffuse optical imaging, phased array imaging, etc. A very new and promising development concerns micro-endoscopy, which allows cellular imaging at the end of a very small optical fiber [98].

Most of these imaging techniques can be used for *Molecular Imaging*, an activity aiming at the *in vivo* characterization and measurement of biological processes at cellular and molecular levels. With optical techniques, molecular imaging makes an extensive use of the fluorescent properties of certain molecules (in particular proteins, e.g. GFP [4]) for imaging of gene expression *in vivo*. With other modalities (like PET, SPECT, MR, CT and even US), molecular imaging can use specific contrast agents or radioactive molecules. For clinical applications, the ultimate goal of molecular imaging is to find the ways to probe much earlier the molecular anomalies that are the basis of a disease rather than to image only its end effects [110].

Some of the recent advances made in Medical Image Analysis could be directly applied (or easily adapted) to Biological Image Analysis. However, the specific nature of biological images (higher resolution, different anatomy and functions, different contrast agents, etc.), requires specific image analysis methods (one can refer to the recent tutorial [103] and to the Mouse Brain Atlas Project [84]). This is particularly true when dealing with *in vivo* microscopic images of cells and vessels.

Our research efforts will be focused to the following generic problems applied to *in vivo* microscopic images:

1. quantitative analysis of microscopic images,
2. detection and quantification of variations in temporal sequences,
3. construction of multiscale representations (from micro to macro).

3.4. Computational Anatomy

The objective of Computational Anatomy (CA) is the modeling and analysis of biological variability of the 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 [5].

[3] This is the case with micro-MRI, Micro-CT, Micro-US devices, and to a less extent with Micro-SPECT and Micro-PET devices.
[4] Green Fluorescent Protein.
[5] 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.

Studying the variability of biological shapes is an old problem (cf. the remarkable book "On Shape and Growth" by D'Arcy Thompson [108]). Significant efforts have been made since that time to develop a theory for statistical shape analysis (one can refer to [90] for a good synthesis, and to the special issue of Neuroimage [107] for recent developments). Despite all these efforts, there is 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 [100]: 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.

3.5. 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 and biology, where CP can be used for instance to better understand the basic processes leading to the apparition 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 [102], [97], [88], [104], [94]). 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 confronting the model with the available biomedical images and signals and possibly also from 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 (e.g. [105], [101]) and the development of new data assimilation methods coping with massive numbers of measurements and unknowns.

There is a hierarchy of modeling levels for CP models of the human body [89]:

- the first level is mainly geometrical, and addresses the construction of a digital description of the anatomy [83], essentially acquired from medical imagery;
- the second level is physical, involving mainly the biomechanical modeling of various tissues, organs, vessels, muscles or bone structures [95];
- the third level is physiological, involving a modeling of the functions of the major biological systems [96] (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 [74].

These different levels of modeling are closely related to each other, and several physiological systems may interact together (e.g. the cardiopulmonary interaction [99]). 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 [96], [97]). 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 biological 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 a 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, ...

3.6. Clinical and Biological 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, this is a necessary condition to see new ideas transformed 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 participate to the stimulation of new ideas and concepts.

ATHENA Project-Team

3. Scientific Foundations

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 [59], Merboldt et al [64] and Taylor et al [70]. 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.

Diffusion Tensor Imaging

Shortly after the first acquisitions of diffusion-weighted images (DWI) were made in vivo [65], [66], 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 [60], [51]. 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 [62], [8] and [61]).

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 [68], [46]. Research groups currently agree that there is complex fiber architecture in most fiber regions of the brain [67]. 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 [72] such as Q-Ball Imaging (QBI) or Diffusion Spectrum Imaging (DSI) [73], [74], [76] 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 [58] or Diffusion Orientation Transform [79]. 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) [76], [73], [74]. 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 [50]. They are at the core of our probabilistic and deterministic tractography algorithms devised to best exploit the full distribution of the fiber ODF (see [48], [3] and [49],[4]).

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 [77], [80] or 4th order Tensor Model [43]. For more details, we refer the reader to our recent article in [53] where we review HOT models and to our article 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.

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 [2], these imaging modalities raise a large amount of mathematical and computational challenges at the core of the research we develop at ATHENA [56].

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 developped 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 [2]. 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. We have also 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 [10], [11]. 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 [10], [11].

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 [55]

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 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 [75] 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. Scientific Foundations

3.1. Formal methods

The study of symbiosis and of biological interactions more in general is the motivation for the work conducted within BAMBOO, but runs in parallel with another important objective. This concerns to (re)visit classical combinatorial (mainly counting / enumerating) and algorithmic problems on strings and (hyper)graphs, and to explore the new variants / original combinatorial and algorithmic problems that are raised by the main areas of application of this project. As the objectives of these formal methods are motivated by biological questions, they are briefly described together with those questions in the next section.

3.2. 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. Experimentaly, 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 pratically, (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.

3.3. 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.

BANG Project-Team

3. Scientific Foundations

3.1. Introduction

The dynamics of complex physical or biophysical phenomena involving many particles, including biological cells - which can be seen as active particles -, can be represented efficiently either by explicitly considering the behaviour of each particle individually or by Partial Differential Equations which, under certain hypotheses, represent local averages over a sufficiently large number of particles.

Since the XIXth century this formalism has shown its efficiency and ability to explain both qualitative and quantitative behaviours. The knowledge that has been gathered on such physical models, on algorithms for solving them on computers, on industrial implementation, opens the hope for success when dealing with life sciences also. This is one of the main goals of BANG. At small spatial scales, or at spatial scales of individual matter components where heterogeneities in the medium occur, agent-based models are developed. They complement the partial differential equation models considered on scales at which averages over the individual components behave sufficiently smoothly.

3.2. Mathematical modelling

What are the relevant physical or biological variables, what are the possible dominant effects ruling their dynamics, how to analyse the information coming out from a mathematical model and interpret them in the real situations under consideration ? These are the questions leading to select a mathematical model, generally also to couple several of them in order to render all physical or biomedical features which are selected by specialist partners (engineers, physicists, medical doctors). These are usually based on Navier-Stokes system for fluids (as in free surface fluid flows), on parabolic-hyperbolic equations (Saint-Venant system for shallow water, flows of electrons/holes in semiconductors, Keller-Segel model of chemotaxis).

3.3. Multiscale analysis

The complete physical or biomedical description is usually complex and requires very small scales. Efficiency of computer resolution leads to simplifications using averages of quantities. Methods allowing to achieve that goal are numerous and mathematically deep. Some examples studied in BANG are

- Coupled multiscale modelling (description of tumours and tissues from the sub-cellular level to the organ scale).
- Description of cell movement from the individual to the collective scales.
- Reduction of full 3d Navier-Stokes system to 2d or 1d hyperbolic equations by a section average (derivation of Saint-Venant system for shallow water).

3.4. Numerical Algorithms

Various numerical methods are used in BANG. They may be based on finite elements or finite volume methods, or stochastic methods for individual agents. Algorithmic improvements are needed in order to take into account the specificity of each model, of their coupling, or their 3D features. Among them we can mention

- Well-balanced schemes for shallow water system.
- Free-surface Navier-Stokes solvers based on a multilayer St-Venant approach.
- Deterministic and stochastic agent based models for the simulation of multi-cellular systems.

3.5. Proliferation dynamics and its control

- Cell division cycle in structured cell populations.
- Physiological and pharmacological control of cell proliferation.
- Physical mechanisms and constraints in cell proliferation.
- Intracellular spatiotemporal dynamics of genes and proteins: p53.
- Cell darwinism and drug resistance in cancer cells.
- Optimisation of cancer chemotherapy.
- Protein polymerisation and application to amyloid diseases.
- Inverse Problem for growth-fragmentation equations.

3.6. Tissue growth, regeneration and cell movements

This research activity aims at studying mathematical models related to tumour development and tissue organisation. Among the many biological aspects, examples are:

- Biomedical aspects of cell-cell interactions at the local and whole organ level.
- Migration of cells in tissues.
- Growth control of living tissues and organs.
- Regenerative medicine.
- Early embryology, and biomechanical aspects of cell interaction.
- Chemotaxis, self-organisation in cell populations.

3.7. Neurosciences

Cortical networks are constituted of a large number of statistically similar neurons in interaction. Each neuron has a nonlinear dynamics and is subject to noise. Moreover, neurological treatment involve several timescales. Multiscale analysis, both in spatial (number of cells) and temporal hence also constitute mathematical foundations of our approaches to neurosciences. In addition to the techniques described in section 3.1 - 3.4, our approach of the activity of large cortical areas involve:

- limit theorems of stochastic interacting particles systems, such as coupling methods or large deviations techniques, as used in mathematical approaches to the statistical physics of gases
- bifurcation analysis of deterministic and stochastic differential equation used to analyze the qualitative behaviour of networks
- singular perturbation theory, geometrical and topological approaches in dynamical systems used to uncover the dynamics in the presence of multiple timescales.

3.8. Free surface flows

Several industrial applications require to solve fluid flows with a free surface. BANG develops algorithms in two directions. Firstly flows in rivers and coastal areas using the Saint-Venant model with applications to dam break and pollution problems in averaged shallow water systems. Secondly, 3D hydrostatic flows by a multilayer Saint-Venant approach and 3D Navier-Stokes flows.

BEAGLE Team

3. Scientific Foundations

3.1. Introduction

As stated above, the research topics of the Beagle Team are centered on the 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 will be 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.We more specifically 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 [47]. 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 [3]. 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 [56]. 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 [4]. 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. To overcome this difficulty, 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 promoterterminator- 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 [64]). As a consequence, this model can be used to study how evolutionary pressures like the ones for robustness or evolvability can shape genome structure [65], [62], [63], [74]. 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 [65], [62], [75]. 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 [55]. 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 recently 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 [49], [48], [50]. 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 [48]. 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 reflexion. 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 [65], and no difference in lifestyles and environment was needed to explain the complexity of the gene regulatory network [48]. 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.

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 [66], [59], 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.

There exist good phylogenic models of punctual mutations on sequences [57], which enable the reconstruction of small parts of ancestral sequences, individual genes for example [67]. But models of whole genome evolution, taking into account large scale events like duplications, insertions, deletions, lateral transfer, rearrangements are just being developped: [77] model punctual mutations as well as duplication and losses of genes, while [52] can reconstruct the evolution of the structure of genomes by inversions. This allows a more comprehensive view of the history of the molecules and the genes, which sometimes have their own historical pattern. But integrative models, considering both nucleotide subsitutions and genome architectures, are still missing.

It is possible to partially reconstruct ancestral genomes for limited cases, by treating separately different types of mutations. It has been done for example for gene content [53], gene order [68], [71], the fate of gene copies after a duplication [61], [45]. All these lead to evolutionary hypotheses on the birth and death of genes [54], on the rearrangements due to duplications [46], [76], on the reasons of variation of genome size [60], [69]. 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 [70], methods for reconstructing the organization of ancestral genomes [72], [51], [73], or for detecting lateral gene transfer events [44], [12]. 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. Scientific Foundations

3.1. Online data analysis

Participants: J-M. Monnez, R. Bar, 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 [61], 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 [58]), which allows to approximate eigenvectors in a stepwise manner. A systematic study of Principal Component and Factorial Analysis has then been leaded by Monnez in the series of papers [64], [62], [63], 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. [59]). 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 [46], [47], we address the following questions:

- Construction of a global test by means of Cramer-von Mises statistic.
- Optimal bandwidth of the kernel used for approximation purposes.

We also obtain sharp estimates on the conditional distribution function in [48].

3.3. Stochastic modeling for complex and biological systems

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 [31] as well as mixed effects for parameter estimation.

• Another important aspect to quantify for PDT 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 leaded us to introduce a new model based on multi-state Markov chains arguments [10], 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 [39]. 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 [65], [68]) 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 [42], 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 [60], but also invoking the Russo-Vallois integration techniques [67]. The specific issue of Volterra equations driven by fBm, which is central for the subdiffusion within proteins problem, is addressed in [43].

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 [38]). Moreover, the fractional Brownian motion cannot be used to model some phenomena for which the regularity varies with time. Hence, many generalization (gaussian or not) of this model has been recently proposed, see for instance [32] for some Gaussian locally self-similar fields, [54] for some non-Gaussian models, [36] for anisotropic models.

Our team has thus contributed [41], [55], [54], [56], [66] and still contributes [35], [37], [36], [57], [49] 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

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 [72] 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 [33] 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. (i) 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 [31], 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.monolix.org) developed 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. (ii) 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 [44]. 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. (iii) 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 [53] 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 [45].

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 [30], [40]. 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 [69], dealing with some very particular families of equations, which do not cover the cases of interest for us.

BIOCORE Project-Team

3. Scientific Foundations

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 the living 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 being irrelevant to the problem at hand, the uncertainties and noise in experiments or even in the biological interactions require the development of 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 the living, 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, be it 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 [7]. 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. The interplay and information transfer between the different levels and scales within a biosystem/ecosystem introduces many new dynamical aspects. 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 to optimize some production or behavior at higher level: for example, control the growth of microalgae via their genetic or metabolic networks, to optimize the production of lipids for bioenergy at the photobioreactor level.

BONSAI Project-Team

3. Scientific Foundations

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 [20], [27], [29], [23], [22]. Members of the team have also a strong expertise in text indexing and compressed index data structures [28], [31], [30]. 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 [32], [26], [25], [24], [17] or non-ribosomal peptides [18]. 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.

3.2. High-performance computing

High-performance computing is another tool that we use to achieve our goals. It covers several paradigms: grids, single-instruction, multiple-data (SIMD) instructions or manycore processors such as graphics cards (GPU). For example, libraries like CUDA and OpenCL also facilitate the use of these manycore processors. These hardware architectures bring promising opportunities for time-consuming bottlenecks arising in bioinformatics.

3.3. 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 contexte, Bayesian models and their MCMC sampling allow to approximate probability distributions over parameters and to describe more biologically relevant models.

CARMEN Team

3. Scientific Foundations

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 [11], [3], [12] 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 [13], [14] and [19] and [2], 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.

CLIME Project-Team

3. Scientific Foundations

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 developed by CEREA: 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 initiatives.

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 value. 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 MétéoFrance, 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).

CORTEX Project-Team

3. Scientific Foundations

3.1. Computational neuroscience

With regards to the progress that has been made in anatomy, neurobiology, physiology, imaging, and behavioral studies, computational neuroscience offers a unique interdisciplinary cooperation between experimental and clinical neuroscientists, physicists, mathematicians and computer scientists. It combines experiments with data analysis and functional models with computer simulation on the basis of strong theoretical concepts and aims at understanding mechanisms that underlie neural processes such as perception, action, learning, memory or cognition.

Today, computational models are able to offer new approaches for the understanding of the complex relations between the structural and the functional level of the brain, thanks to models built at several levels of description. In very precise models, a neuron can be divided in several compartments and its dynamics can be described by a system of differential equations. The spiking neuron approach (*cf.* § 3.2) proposes to define simpler models concentrated on the prediction of the most important events for neurons, the emission of spikes. This allows to compute networks of neurons and to study the neural code with event-driven computations.

Larger neuronal systems are considered when the unit of computation is defined at the level of the population of neurons and when rate coding and/or correlations are supposed to bring enough information. Studying Dynamic Neural Fields (*cf.* § 3.3) consequently lays emphasis on information flows between populations of neurons (feed-forward, feed-back, lateral connectivity) and is well adapted to defining high-level behavioral capabilities related for example to visuomotor coordination.

Furthermore, these computational models and methods have strong implications for other sciences (e.g. computer science, cognitive science, neuroscience) and applications (e.g. robots, cognitive prosthesis) as well (*cf.* § 4.1). In computer science, they promote original modes of distributed computation (*cf.* § 3.5); in cognitive science, they have to be related to current theories of cognition (*cf.* § 3.6); in neuroscience, their predictions have to be related to observed behaviors and measured brain signals (*cf.* § 3.4).

3.2. Computational neuroscience at the microscopic level: spiking neurons and networks

Computational neuroscience is also interested in having more precise and realistic models of the neuron and especially of its dynamics. We consider that the latter aspect cannot be treated at the single unit level only; it is also necessary to consider interactions between neurons at the microscopic scale.

On one hand, compartmental models describe the neuron at the inner scale, through various compartments (axon, synapse, cellular body) and coupled differential equations, allowing to numerically predict the neural activity at a high degree of accuracy. This, however, is intractable if analytic properties are to be derived, or if neural assemblies are considered. We thus focus on phenomenological punctual models of spiking neurons, in order to capture the dynamic behavior of the neuron isolated or inside a network. Generalized conductance based leaky integrate and fire neurons (emitting action potential, i.e. spike, from input integration) or simplified instantiations are considered in our group.

On the other hand, one central issue is to better understand the precise nature of the neural code. From rate coding (the classical assumption that information is mainly conveyed by the firing frequency of neurons) to less explored assumptions such as high-order statistics, time coding (the idea that information is encoded in the firing time of neurons) or synchronization aspects. At the biological level, a fundamental example is the synchronization of neural activities, which seems to play a role in, e.g., olfactory perception: it has been observed that abolishing synchronization suppresses the odor discrimination capability. At the computational

level, recent theoretical results show that the neural code is embedded in periodic firing patterns, while, more generally, we focus on tractable mathematical analysis methods coming from the theory of nonlinear dynamical systems.

For both biological simulations and computer science emerging paradigms, the rigorous simulation of large neural assemblies is a central issue. Our group is at the origin, up to our best knowledge, of the most efficient event-based neural network simulator (Mvaspike), based on well-founded discrete event dynamic systems theory, and now extended to other simulation paradigms, thus offering the capability to push the state of the art on this topic.

3.3. Computational neuroscience at the mesoscopic level: dynamic neural field

Our research activities in the domain of computational neurosciences are also interested in the understanding of higher brain functions using both computational models and robotics. These models are grounded on a computational paradigm that is directly inspired by several brain studies converging on a distributed, asynchronous, numerical and adaptive processing of information and the continuum neural field theory (CNFT) provides the theoretical framework to design models of population of neurons.

This mesoscopic approach underlines the fact that the number of neurons is very high, even in a small part of tissue, and proposes to study neuronal models in a continuum limit where space is continuous and main variables correspond to synaptic activity or firing rates in population of neurons. This formalism is particularly interesting because the dynamic behavior of a large piece of neuronal tissue can be studied with differential equations that can integrate spatial (lateral connectivity) and temporal (speed of propagation) characteristics and display such interesting behavior as pattern formation, travelling waves, bumps, etc.

The main cognitive tasks we are currently interested in are related to sensorimotor systems in interaction with the environment (perception, coordination, planning). The corresponding neuronal structures we are modeling are part of the cortex (perceptive, associative, frontal maps) and the limbic system (hippocampus, amygdala, basal ganglia). Corresponding models of these neuronal structures are defined at the level of the population of neurons and functioning and learning rules are built from neuroscience data to emulate the corresponding information processing (filtering in perceptive maps, multimodal association in associative maps, temporal organization of behavior in frontal maps, episodic memory in hippocampus, emotional conditioning in amygdala, selection of action in basal ganglia). Our aim is to iteratively refine these models, implement them on autonomous robots and make them cooperate and exchange information, toward a completely adaptive, integrated and autonomous behavior.

3.4. Brain Signal Processing

The observation of brain activity and its analysis with appropriate data analysis techniques allow to extract properties of underlying neural activity and to better understand high level functions. This study needs to investigate and integrate, in a single trial, information spread in several cortical areas and available at different scales (MUA, LFP, ECoG, EEG).

One major problem is how to be able to deal with the variability between trials. Thus, it is necessary to develop robust techniques based on stable features. Specific modeling techniques should be able to extract features investigating the time domain and the frequency domain. In the time domain, template-based unsupervised models allows to extract graphic-elements. Both the average technique to obtain the templates and the distance used to match the signal with the templates are important, even when the signal has a strong distorted shape. The study of spike synchrony is also an important challenge. In the frequency domain, features such as phases, frequency bands and amplitudes contain different pieces of information that should be properly identified using variable selection techniques. In both cases, compression techniques such as PCA or ICA can reduce the fluctuations of the cortical signal. Then, the designed models have to be able to track the dynamic evolution of these features over the time.

Another problem is how to integrate information spreading in different areas and relate this information in a proper time window of synchronization to behavior. For example, feedbacks are known to be very important to better understand the closed-loop control of a hand grasping movement. However, from the preparatory signal and the execution of the movement to the visual and somatosensory feedbacks, there is a delay. It is thus necessary to use stable features to build a mapping between areas using supervised models taking into account a time window shift.

Several recoding techniques are taken into account, providing different kinds of information. Some of them provide very local information such as multiunit activities (MUA) and local field potential (LFP) in one or several well-chosen cortical areas. Other ones provide global information about close regions such as electrocorticography (ECoG) or the whole scalp such as electroencephalography (EEG). If surface electrodes allow to easily obtain brain imaging, it is more and more necessary to better investigate the neural code.

3.5. Connectionist parallelism

Connectionist models, such as neural networks, are among the first models of parallel computing. Artificial neural networks now stand as a possible alternative with respect to the standard computing model of current computers. The computing power of these connectionist models is based on their distributed properties: a very fine-grain massive parallelism with densely interconnected computation units.

The connectionist paradigm is the foundation of the robust, adaptive, embeddable and autonomous processings that we aim at developing in our team. Therefore their specific massive parallelism has to be fully exploited. Furthermore, we use this intrinsic parallelism as a guideline to develop new models and algorithms for which parallel implementations are naturally made easier.

Our approach claims that the parallelism of connectionist models makes them able to deal with strong implementation and application constraints. This claim is based on both theoretical and practical properties of neural networks. It is related to a very fine parallelism grain that fits parallel hardware devices, as well as to the emergence of very large reconfigurable systems that become able to handle both adaptability and massive parallelism of neural networks. More particularly, digital reconfigurable circuits (e.g. FPGA, Field Programmable Gate Arrays) stand as the most suitable and flexible device for low cost fully parallel implementations of neural models, according to numerous recent studies in the connectionist community. We carry out various arithmetical and topological studies that are required by the implementation of several neural models onto FPGAs, as well as the definition of hardware-targetted neural models of parallel computation.

This research field has evolved within our team by merging with our activities in behavioral computational neuroscience. Taking advantage of the ability of the neural paradigm to cope with strong constraints, as well as taking advantage of the highly complex cognitive tasks that our behavioral models may perform, a new research line has emerged that aims at defining a specific kind of brain-inspired hardware based on modular and extensive resources that are capable of self-organization and self-recruitment through learning when they are assembled within a perception-action loop.

3.6. The embodiment of cognition

Recent theories from cognitive science stress that human cognition emerges from the interactions of the body with the surrounding world. Through motor actions, the body can orient toward objects to better perceive and analyze them. The analysis is performed on the basis of physical measurements and more or less elaborated emotional reactions of the body, generated by the stimuli. This elicits other orientation activities of the body (approach and grasping or avoidance). This elementary behavior is made possible by the capacity, at the cerebral level, to coordinate the perceptive representation of the outer world (including the perception of the body itself) with the behavioral repertoire that it generates either on the physical body (external actions) or on a more internal aspect (emotions, motivations, decisions). In both cases, this capacity of coordination is acquired from experience and interaction with the environment.

The theory of the situatedness of cognition proposes to minimize representational contents (opposite to complex and hierarchical representations) and privileges simple strategies, more directly coupling perception and action and more efficient to react quickly in the changing environment.

A key aspect of this theory of intelligence is the Gibsonian notion of affordance: perception is not a passive process and, depending on the current task, objects are discriminated as possible "tools" that could be used to interact and act in the environment. Whereas a scene full of details can be memorized in very different and costly ways, a task-dependent description is a very economical way that implies minimal storage requirements. Hence, remembering becomes a constructive process.

For example with such a strategy, the organism can keep track of relevant visual targets in the environment by only storing the movement of the eye necessary to foveate them. We do not memorize details of the objects but we know which eye movement to perform to get them: The world itself is considered as an external memory.

Our agreement to this theory has several implications for our methodology of work. In this view, learning emerges from sensorimotor loops and a real body interacting with a real environment are important characteristics for a learning protocol. Also, in this view, the quality of memory (a flexible representation) is preferred to the quantity of memory.

DEMAR Project-Team

3. Scientific Foundations

3.1. Modelling and identification of the sensory-motor system

Participants: Mitsuhiro Hayashibe, Christine Azevedo Coste, David Guiraud, Philippe Poignet.

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 contribution 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 to 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 teams 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, Guy Cathébras, Fabien Soulier, Serge Bernard.

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 system 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. Scientific Foundations

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).

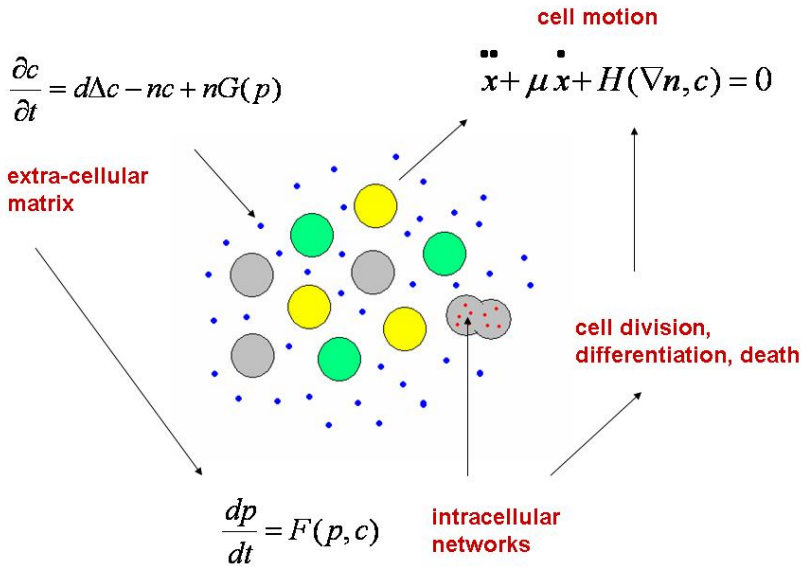


Figure 1. Schema of multi-scale models of cell dynamics: DPD-PDE-ODE models.

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 [35] 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 [36] 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 ([41], [43]), and they proved effective in hematopoiesis modelling ([30], [31], [33]). 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 [43]. The nonlocality can be either in the structure variable or in the time variable [30]. 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 [37].

3.5. Stochastic Equations

How identical cells perform different tasks may depend on deterministic factors, like external signals or pre-programming, or on stochastic factors. Intra-cellular processes are inherently noisy due to low numbers of molecules, complex interactions, limited number of DNA binding sites, the dynamical nature of molecular interactions, etc. Yet at the population level, deterministic and stochastic systems can behave the same way because of averaging over the entire population. This is why it is important to understand the causes and the roles of stochasticity in intra-cellular processes. In its simplest form, stochastic modelling of gene regulation networks considers the evolution of a low number of molecules (integer number) as they are synthesized, bound to other molecules, or degraded. The number n(t) of molecules at time t is a stochastic process whose probability transition to n+1 or n-1 is governed by a specific law. In some cases, master equations can yield analytical solutions for the probability distribution of n, P(n(t)). Numerically, efficient algorithms have been developed (Gillespie algorithms and variants) to handle statistically exact solutions of biochemical reactions. Recently, these algorithms have been adapted to take into account time delays. This allows a stochastic description of delayed regulatory feedback loops, both at the intra-cellular and the population levels. Another approach with stochastic differential equation, using Langevin equations is relevant to study extrinsic sources of noise on a system. A thesis (R. Yvinec) supervised by L. Pujo-Menjouet and M.C. Mackey devoted to "stochastic differential equations", started in Lyon on October 2009.

DYLISS Team

3. Scientific Foundations

3.1. Knowledge representation with constraint programming

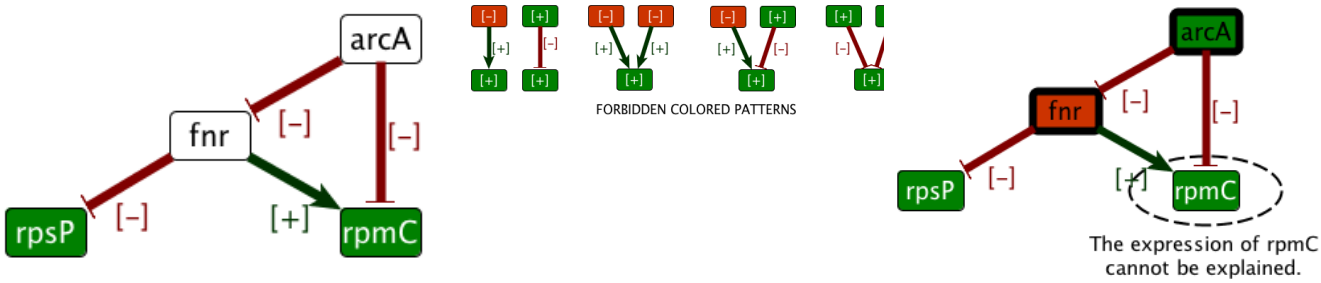
Biological networks are built with data-driven approaches aiming at translating genomic information into a functional map. Most methods are based on a probabilistic framework which defines a probability distribution over the set of models. The reconstructed network is then defined as the most likely 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. 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 optimization problem[6].

Solving these computational issues requires addressing NP-hard qualitative (non-temporal) issues, based on a notion of causality. We have developed a long-term collaboration with Potsdam University in order to use a logical paradigm named **Answer Set Programming** [27], [30] to solve these optimization issues. Applied on transcriptomic or cancer networks, our methods identified which regions of a large-scale network shall be corrected [1], and proposed robust corrections [5]. 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 optimization issues explored in systems biology.

Using these technologies requires to revisit and reformulate optimization problems at hand in order both to decrease the search space size in the grounding part of the process and to optimize 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.

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 [2]. 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 [29]. This property can be derived into constraints to describe the set of admissible



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).

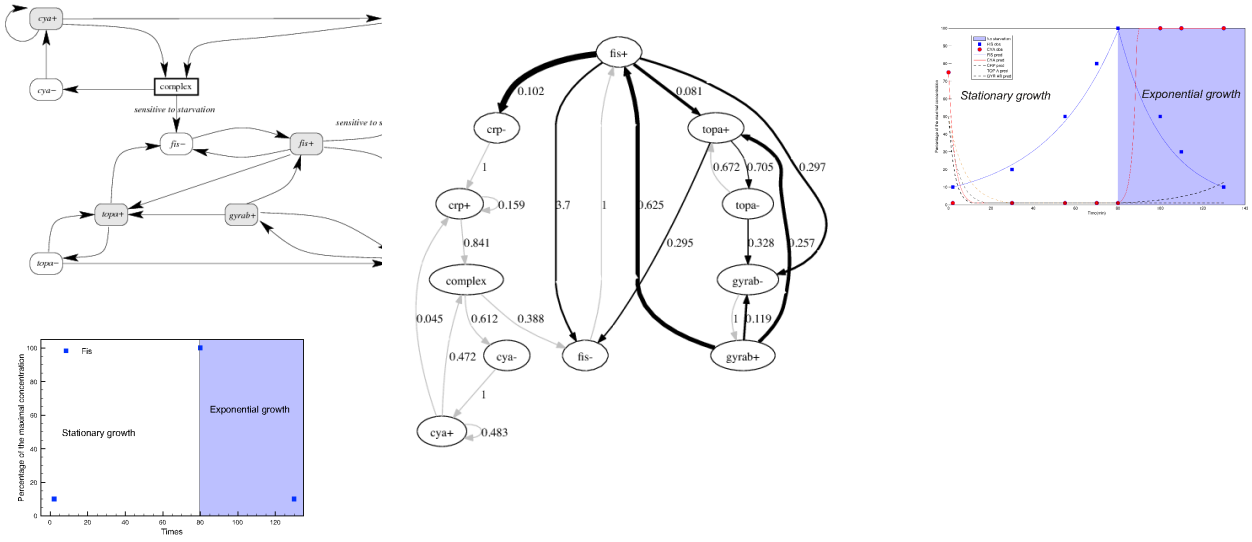
Step 4 . Excerpt from the ASP program.

```
vertex(fnr).                          1{labelV(I ,+;-)}1 :- vertex(I).        receive(I,+) :- labelE(J,I,S), labelV(J,S).
edge(fnr,rpsP).                       labelV(I ,S) :- observedV(I,S).         receive(I,-) :- labelE(J,I,S), labelV(J,T), S≠T.

observedE(fnr,rpsP,-).                1{labelE(J,I,+;-)}1 :- edge(J,I).       :- labelV (I,S), not receive(I,S).
observedV(rpsP,-).                    labelE(J,I,S) :- observedE(J,I,S).
```

Figure 1. Excerpt from the ASP program identifying which expression of non-observed nodes (white nodes) are fixed by partial observations and rules derived from the system dynamics. The 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 [6], [5] .

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. We now plan to apply these techniques to reduced networks representing the main pathways and actors automatically generated from the integrative methods developed in Axis 1. 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.



Input data . Qualitative description of the system dynamics at the transcription level (interaction graph) and 3 concentration measurements of the fis protein (population scale).

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.

Prediction of the Cya protein concentration (red curve) fits with observations. Additionally, literature evidences that high sensitivity ETG transitions correspond to key interaction in E. Coli response to nutritional stress.

Figure 2. Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems. Identification of key interactions [2].

3.3. Grammatical inference and highly expressive structures

Our main field of expertise in **machine learning** concerns grammatical models with a long -term know-how 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 [4], leading to a tool named Protomata-learner. As an example, this tools allows to properly model the multi-domain function of the protein family TNF, which is impossible with other existing probabilistic-based approach (see Fig. 3). Our future goal is to demonstrate the relevance of formal language theory 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.

Moving forward to context-free grammars instead of regular patterns increases **parsing** complexity. 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 [31]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [34], [33]. 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 [9]. 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.

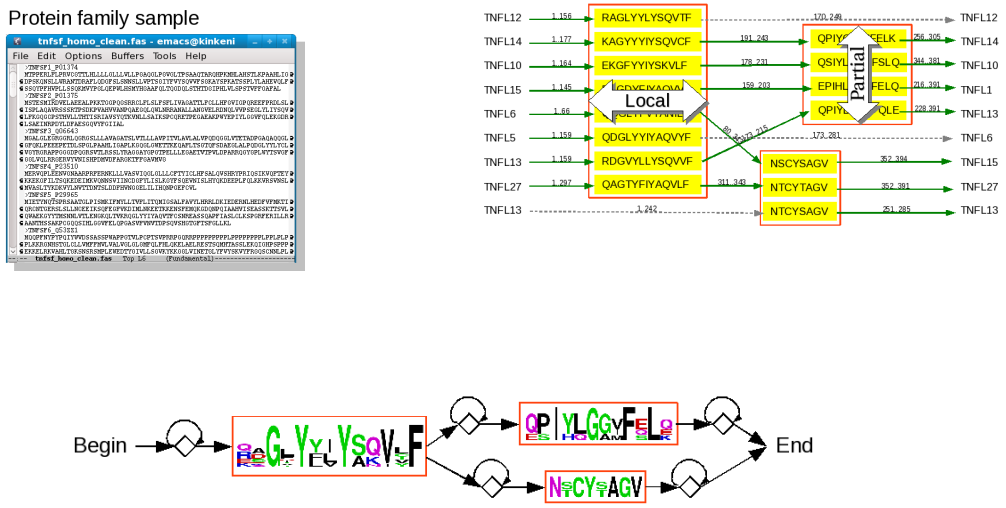


Figure 3. ***Protomata Learner workflow****. Starting from a set of protein sequences (up left), a partial local alignment is computed (up right) and an automaton is inferred, which models the family and allows to search for its unknown members (down) .*

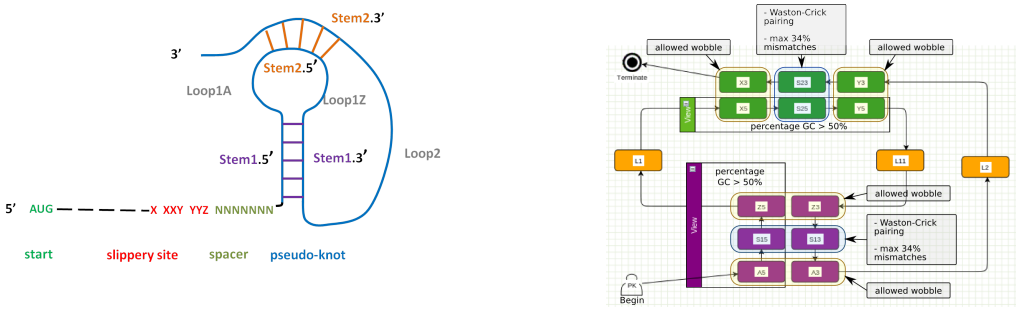


Figure 4. ***Left****: A typical RNA structure: the pseudo-knot.* ***Right****: graphical modeling of a pseudo-knot with String Variable Grammars used in the* ***Logol framework****.*

FLUMINANCE Project-Team

3. Scientific Foundations

3.1. Fluid flow analysis and modeling

Turbulent fluid flows involved in environmental or industrial applications are complex. In fluid mechanics laboratories, canonical turbulent shear flows have been studied for many years and a relatively clear picture of their underlying structure exists. However, the direct applicability of these efforts to real relevant flows, which often occur in complex geometries and in the presence of multiple non canonical influences, like cross-shear, span wise non-uniform and thermal stratification, is still unknown. In addition, the turbulence can be characterized by Reynolds number ranging between 10^3 and 10^4, corresponding to transitional regime for which the use of classical turbulence models is limited.

In this context, we have performed research studies on turbulent shear flows of low velocities by tackling crucial topics of measurements, analysis and modeling of environmental and industrial flows in presence of non-canonical influences. This concerns more precisely the study of the interaction between a mixing layer and circular cylinder wake flow, the study of wake flow with span wise non uniformity, the study of mixing layer under the influence of thermal stratification and the study of mixing layer forced between non-uniform flows. The analysis of these flows has required the design of adequate dynamical models, using proper orthogonal decomposition and Galerkin projection. Understanding issues such as the mechanisms of heat and mass transfer involved in these shear flows provides meaningful information for the control of relevant engineering flows and the design of new technologies. To investigate more thoroughly these complex flows numerical and experimental tools have been designed. An immersed boundary method was proposed to mimic complex geometries into Direct Numerical simulation (DNS) and Large Eddy Simulations (LES) codes. A novel anemometer has been designed and implemented for the simultaneous measurement of velocity and temperature in air flows with a single hot-wire probe.

Mixing layer wake interaction
We have investigated the vortex shedding of a circular cylinder immersed in a plane turbulent mixing layer. For a centre span Reynolds number of 7500, the wake flow splits into three regions: a high-velocity wake, a low-velocity wake and a region of interaction in the middle span of the body. A strong unsteady secondary flow is observed, and explained with span wise base pressure gradients. Unexpected features are found for formation length and the base pressure along the span of the cylinder. In the high-velocity side, where the local Reynolds number is the highest, the formation length is longest. Based on the formation length measurements it was shown that as a function of the centre span Reynolds number, the wake flows behaves as circular cylinder in uniform flow. Three cells with a constant frequency with adjacent dislocations are observed. For each cell, a shedding mode was suggested. The relation of the secondary flow to the frequencies was examined. All the observations were analyzed by analogical reasoning with other flows. This pointed out the action of the secondary flow in the high-velocity side regarded as a wake interference mechanism.

Low order complex flow modeling
We have proposed improvements to the construction of low order dynamical systems (LODS) for incompressible turbulent external flows. The reduced model is obtained by means of a Proper Orthogonal Decomposition (POD) basis extracted through a truncated singular value decomposition of the flow auto-correlation matrix built from noisy PIV experimental velocity measurements. The POD modes are then used to formulate a reduced dynamical system that contains the main features of the flow. This low order dynamical system (LODS) is obtained through a Galerkin projection of the Navier-Stokes Equations on the POD basis. Usually, the resulting system of ordinary differential equations presents stability problems due to modes truncation and numerical uncertainties, especially when working on experimental data. The technique we proposed relies on an optimal control approach to estimate the dynamical system coefficients and its initial condition. This allows us to recover a reliable and stable spatio-temporal reconstruction of the large scales of the flow. The technique

has been assessed on the near wake behind a cylinder observed through very noisy PIV measurement. It has been also evaluated for configurations involving a rotating cylinder.

Studies on complex 3D dynamical behavior resulting from the interaction between a plane mixing layer and the wake of a cylinder have been also investigated using POD representation, applied to data from two synchronized 2D PIV systems (Dual-plane PIV). This approach allowed us to construct a 3D-POD representation. An analysis of the correlations shows different length scales in the regions dominated by wake like structures and shear layer type structures [2]. In order to characterize the particular organization in the plane of symmetry, a Galerkin projection from a slice POD has been performed. This led to a low-dimensional dynamical system that allowed the analysis of the relationship between the dominant frequencies. This study led to a reconstruction of the dominant periodic motion suspected from previous studies [41]. This work allowed us to make a link between the three-dimensional organization and the secondary unsteady motion from the low velocity side to the high velocity side of the mixing layer, appearing in this highly 3D flow configuration.

Direct and Large Eddy simulations of complex flows
We have proposed a direct forcing method better suited to the use of compact finite difference schemes in Direct Numerical Simulation. The new forcing creates inside the body an artificial flow preserving the no-slip condition at the surface but reducing the step-like change of the velocity derivatives across the immersed boundary. This modification led to improve results both qualitatively and quantitatively for conventional and complex flow geometries [50].

Three-dimensional direct numerical simulations have been performed for vortex shedding behind cylinders. We focused in particular on cases for which the body diameter and the incoming flow involved span wise linear non-uniformity. Four configurations were considered: the shear flow, the tapered cylinder and their combinations, which gave rise namely to the adverse and aiding cases. In contrast with the observations of other investigators, these computations highlighted distinct vortical features between the shear case and the tapered case. In addition, it was observed that the shear case and the adverse case (respectively the tapered and aiding case), yielded similarities in flow topology. This phenomenon was explained by the span wise variations of the ratio of mean velocity and the cylinder diameter which seemed to govern these flows. Indeed, it was observed that large span wise variations of U/D seemed to enhance three-dimensionality, through the appearance of vortex-adhesions and dislocations. Span wise cellular pattern of vortex shedding were identified. Their modifications in cell size, junction position and number were correlated with the variation of U/D. In the Lee side of the obstacle a wavy secondary motion was identified. Induced secondary flow due to the bending of Karman vortices in the vicinity of vortex-adhesion and dislocations was suggested to explain this result [49].

LES and experimental wake flow database
We contributed to the study of flow over a circular cylinder at Reynolds number $Re = 3900$. Although this classical flow is widely documented in the literature, especially for this precise Reynolds number, which leads to a sub critical flow regime, there is no consensus about the turbulence statistics immediately just behind the obstacle. This flow has been studied both numerically with Large Eddy Simulation and experimentally with Hot-Wire Anemometry and Particle Image Velocimetry. The numerical simulation has been performed using high-order schemes and the specific Immersed Boundary Method previously mentioned. We focused on turbulence statistics and power spectra in the near wake up to 10 diameters. Statistical estimation is shown to need large integration times increasing the computational cost and leading to an uncertainty of about 10% for most flow characteristics considered in this study. The present numerical and experimental results are found to be in good agreement with previous Large Eddy Simulation data. Our study has exhibited significant differences compared with the experimental data found in the literature. The obtained results attenuate previous numerical-experimental controversy for this type of flows [11].

Simultaneous velocity temperature measurements in turbulent flows

We have worked on the design of a novel anemometer for the simultaneous measurement of velocity and temperature in airflows with a single hot wire probe. The principle of periodically varying the overheat ratio of the wire has been selected and applied through a tunable electronic chain. Specific methods were developed for the calibration procedure and the signal processing. The accuracy of the measurements was assessed by means of Monte-Carlo simulations. Accurate results were provided for two types of turbulent non-isothermal flows, a coaxial heated jet and a low speed thermal mixing. The particular interest of the synchronization of the two measurements has been emphasized during the PhD thesis of T. Ndoye.

A new dynamic calibration technique has been developed for hot-wire probes. The technique permits, in a short time range, the combined calibration of velocity, temperature and direction calibration of single and multiple hot-wire probes. The calibration and measurements uncertainties were modeled, simulated and controlled, in order to reduce their estimated values.

3.2. Fluid motion analysis

Flow visualization has been a powerful tool to depict or to understand flow feature properties. Efforts to develop high-quality flow visualization techniques date back over a century. The analysis of the recorded images consisted firstly to a qualitative interpretation of the streak lines leading to an overall global insight into the flow properties but lacking quantitative details on important parameters such as velocity fields or turbulence intensities. Point measurement tools such as hot wire probes or Laser Doppler Velocimetry have typically provided these details. As these probes give information only at the point where they are placed, simultaneous evaluations at different points require to dispose a very large number of probes and the evaluation of unsteady field (most of the flows are unsteady) is almost unachievable with them.

In an effort to avoid the limitations of these probes, the Particle Image Velocimetry (PIV), a non-intrusive diagnostic technique, has been developed in the last two decades [40]. The PIV technique enables obtaining velocity fields by seeding the flow with particles (e.g. dye, smoke, particles) and observing the motion of these tracers. In computer vision, the estimation of the projection of the apparent motion of a 3D scene onto the image plane, refereed 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 [45]. 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 the kinematical and dynamical laws governing the observed fluid flows. Besides, within this framework it is also much easier to define characteristic features estimation 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. This route has demonstrated to be much more robust to scalar image. Several studies in this vein have strengthened our skills in this domain. All the following approaches have been either formulated within a statistical Markov Random Fields modeling or either within a variational framework. For a thorough description of these approaches see [7].

ICE data model and div-curl regularization This fluid motion estimator is constructed on a data model derived from the Integration of the Continuity Equation (ICE data model) [5] and includes a second order regularization scheme enabling to preserve blobs of divergence and curl. Intensive evaluations of this estimator on flow prototypes mastered in laboratory have shown that this estimator led to the same order of accuracy as the best PIV techniques but for an increase information density. This ability to get dense flow fields allowed us estimating proper vorticity or divergence maps without resorting to additional post-processing interpolation schemes.

Schlieren Image velocimetry We have addressed the problem of estimating the motion of fluid flows visualized with the Schlieren technique. Such an experimental visualization system is well known in fluid mechanics and it enables the visualization of unseeded flows. This technique authorizes the capture of phenomena that are impossible to visualize with particle seeding such as natural convection, phonation flow, breath flow and allows the setting of large scale experiments. Since the resulting images exhibit very low intensity contrasts, classical motion estimation methods based on the brightness constancy assumption (correlation-based approaches, optical flow methods) are completely inefficient. The global energy function we have defined for Schlieren images is composed of i) a specific data model accounting for the fact that the observed luminance is related to the gradient of the fluid density, and ii) a specific constrained div-curl regularization term. To date there exists no motion estimator allowing estimating accurately dense velocity fields on Schlieren images.

Low order fluid motion estimator This low-dimensional fluid motion estimator [6] is based on the Helmholtz decomposition, which consists in representing the velocity field as the sum of a divergence-free component and a curl-free one. In order to provide a low-dimensional solution, both components have been approximated using a discretization of the vorticity (curl of the velocity vector) and divergence maps through regularized Dirac measures [44]. The resulting so-called irrotational (resp. solenoidal) field is then represented by a linear combination of basis functions obtained by a convolution product of the Green kernel gradient and the vorticity map (resp. the divergence map). The coefficient values and the basis function parameters are obtained by minimizing a function formed by an integrated version of the mass conservation principle of fluid mechanics.

Potential functions estimation and finite mimetic differences We have studied a direct estimation approach of the flow potential functions (respectively the *stream* function and the *velocity* potential) from two consecutive images. The estimation has been defined on the basis of a high order regularization scheme and has been implemented through mimetic difference methods[12]. With these approaches the discretization preserves basic relationships of continuous vector analysis. Compared to previous discretization scheme based on auxiliary div-curl variables, the considered technique appeared to be numerically much more stable and led to an improve accuracy.

2D and 3D atmospheric motion layer estimation In this study, we have explored the problem of estimating mesoscales dynamics of atmospheric layers from satellite image sequences. Due to the intrinsic sparse 3-dimensional nature of clouds and to large occluded zones caused by the successive overlapping of cloud layers, the estimation of accurate layered dense motion fields is an intricate issue. Relying on a physically sound vertical decomposition of the atmosphere into layers, we have proposed two dense motion estimators for the extraction of multi-layer horizontal (2D) and 3D wind fields. These estimators are expressed as the minimization of a global function that includes a data-driven term and a spatio-temporal smoothness term. A robust data term relying on shallow-water mass conservation model has been proposed to fit sparse observations related to each layer. In the 3D case, the layers are interconnected through a term modeling mass exchanges at the layers surfaces frontiers [9].

A novel spatio-temporal regularizer derived from the shallow-water momentum conservation model has been considered to enforce temporal consistency of the solution along time. These constraints are combined with a robust second-order regularizer preserving divergent and vorticity structures of the flow. Besides, a two-level motion estimation scheme has been settled to overcome the limitations of the multiresolution incremental estimation scheme when capturing the dynamics of fine mesoscale structures. This alternative approach relies on the combination of correlation and optical-flow observations. An exhaustive evaluation of the novel method has been first performed on a scalar image sequence generated by Direct Numerical Simulation of a turbulent bi-dimensional flow. Based on qualitative experimental comparisons, the method has also been assessed on a Meteosat infrared image sequence.

3.3. Data assimilation and Tracking of characteristic fluid features

Classical motion estimation techniques usually proceed on pairs of two successive images, and do not enforce temporal consistency. This often induces an estimation drift which is essentially due to the fact that motion estimation is formulated as a local process in time. No adequate physical dynamics law, or conservation law,

related to the observed flow, is taken into account over long time intervals by the usual motion estimators. The estimation of an unknown state variable trajectory on the basis of specified dynamical laws and some incomplete and noisy measurements of the variable of interest can be either conducted through optimal control techniques or through stochastic filtering approach. These two frameworks have their own advantages and deficiencies. We rely indifferently on both approaches.

Stochastic filtering for fluid motion tracking We have proposed a recursive Bayesian filter for tracking velocity fields of fluid flows. The filter combines an Îto diffusion process associated to 2D vorticity-velocity formulation of Navier-Stokes equation and discrete image error reconstruction measurements. In contrast to usual filters, designed for visual tracking problem, our filter combines a continuous law for the description of the vorticity evolution with discrete image measurements. We resort to a Monte-Carlo approximation based on particle filtering. The designed tracker provides a robust and consistent estimation of instantaneous motion fields along the whole image sequence. In order to handle a state space of reasonable dimension for the stochastic filtering problem, the motion field is represented as a combination of adapted basis functions. The basis functions are derived from a mollification of Biot-Savart integral and a discretization of the vorticity and divergence maps of the fluid vector field. The output of such a tracking is a set of motion fields along the whole time range of the image sequence. As the time discretization is much finer than the frame rate, the method provides consistent motion interpolation between consecutive frames. In order to reduce further the dimensionality of the associated state space when we are facing a large number of motion basis functions, we have explored a new dimensional reduction approach based on dynamical systems theory. The study of the stable and unstable directions of the continuous dynamics enables to construct an adaptive dimension reduction procedure. It consists in sampling only in the unstable directions, while the stable ones are treated deterministically [6].

When the likelihood of the measurement can be modeled as Gaussian law, we have also investigated the use of so-called ensemble Kalman filtering for fluid tracking problems. This kind of filters introduced for the analysis of geophysical fluids is based on the Kalman filter update equation. Nevertheless, unlike traditional Kalman filtering setting, the covariances of the estimation errors, required to compute the Kalman gain, rely on an ensemble of forecasts. Such a process gives rise to a Monte Carlo approximation for a family of stochastic non linear filters enabling to handle state spaces of large dimension. We have recently proposed an extension of this technique that combines sequential importance sampling and the propagation law of ensemble kalman filter. This technique leads to ensemble Kalman filter with an improve efficiency. This appears to be a generalization of the optimal importance sampling strategy we proposed in the context of partial conditional Gaussian trackers [1].

Variational assimilation technique

We investigated the use of variational framework for the tracking from image sequence of features belonging to high dimensional spaces. This framework relies on optimal control principles as developed in environmental sciences to analyze geophysical flows [46], [47]. Within the PhD of Nicolas Papadakis [10], we have first devised a data assimilation technique for the tracking of closed curves and their associated motion fields. The proposed approach enables a continuous tracking along an image sequence of both a deformable curve and its associated velocity field. Such an approach has been formalized through the minimization of a global spatio-temporal continuous cost functional, with respect to a set of variables representing the curve and its related motion field. The resulting minimization sequence consists in a forward integration of an evolution law followed by a backward integration of an adjoint evolution model. The latter pde includes a term related to the discrepancy between the state variables evolution law and discrete noisy measurements of the system. The closed curves are represented through implicit surface modeling [48], whereas the motion is described either by a vector field or through vorticity and divergence maps according to the type of targeted application. The efficiency of the approach has been demonstrated on two types of image sequences showing deformable objects and fluid motions.

More recently assimilation technique for the direct estimation of atmospheric wind field from pressure images have been proposed [4]. These techniques rely on a brightness variation model of the intensity function. They do not include anymore motion measurements provided by external motion estimators. The resulting estimator

allows us to recover accurate fluid motion fields and enables tracking dense vorticity maps along an image sequence.

3.4. Visual servoing

Nowadays, visual servoing is a widely used technique is robot control. It consists in using data provided by a vision sensor for controlling the motions of a robot [43]. Various sensors can be considered such as perspective cameras, omnidirectional cameras, 2D ultrasound probes or even virtual cameras. In fact, this technique is historically embedded in the larger domain of sensor-based control [51] so that other sensors than vision sensors can be properly used. On the other hand, this approach was first dedicated to robot arms control. Today, much more complex system can be considered like humanoid robots, cars, submarines, airships, helicopters, aircrafts. Therefore, visual servoing is now seen as a powerful approach to control the state of dynamic systems.

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 the image. A control law is then designed so that these visual features reach a desired value $\mathbf{s}^*$ related to the desired state of the system. The control principle is thus to regulate 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 camera instantaneous velocity $\mathbf{v}$

$$\dot{\mathbf{s}} = \mathbf{L_s}\,\mathbf{v} + \frac{\partial \mathbf{s}}{\partial t} \tag{87}$$

where the term $\frac{\partial \mathbf{s}}{\partial t}$ describes the non-stationary behavior of $\mathbf{s}$. Typically, if we try to ensure an exponential decoupled decrease of the error signal and if we consider the camera velocity as the input of the robot controller, the control law writes as follow

$$\mathbf{v} = -\lambda \widehat{\mathbf{L_s}}^{+}\mathbf{e} - \widehat{\mathbf{L_s}}^{+}\widehat{\frac{\partial \mathbf{e}}{\partial t}} \tag{88}$$

with λ a proportional gain that has to be tuned to minimize the time-to-convergence, $\widehat{\mathbf{L_s}}^{+}$ the pseudo-inverse of a model or an approximation of $\mathbf{L_s}$ and $\widehat{\frac{\partial \mathbf{e}}{\partial t}}$ an estimation of $\frac{\partial \mathbf{e}}{\partial t}$.

The behavior of the closed-loop system is then obtained, from (2), by expressing the time variation of the error $\mathbf{e}$

$$\dot{\mathbf{e}} = -\lambda \mathbf{L_s}\widehat{\mathbf{L_s}}^{+}\mathbf{e} - \mathbf{L_s}\widehat{\mathbf{L_e}}^{+}\widehat{\frac{\partial \mathbf{e}}{\partial t}} + \frac{\partial \mathbf{e}}{\partial t}. \tag{89}$$

As can be seen, visual servoing explicitly relies on the choice of the visual features $\mathbf{s}$ and then on the related interaction matrix; that is the key point of this approach. Indeed, this choice must be performed very carefully. Especially, an isomorphism between the camera pose and the visual features is required to ensure that the convergence of the control law will lead to the desired state of the system. An optimal choice would result in finding visual features leading to a diagonal and constant interaction matrix and, consequently, to a linear decoupled system for which the control problem is well known. Thereafter, the isomorphism as well as the global stability would be guaranteed. In addition, since the interaction matrix would present no more nonlinearities, a suitable robot trajectory would be ensured.

However, finding such visual features is a very complex problem and it is still an open issue. Basically, this problem consists in building the visual features $\mathbf{s}$ from the nonlinear visual measurements $\mathbf{m}$ so that the interaction matrix related to $\mathbf{s}$ becomes diagonal and constant or, at least, as simple as possible.

On the other hand, a robust extraction, matching (between the initial and desired measurements) and real-time spatio-temporal tracking (between successive measurements) have to be ensure but have proved to be a complex task, as testified by the abundant literature on the subject. Nevertheless, this image process is, to date, a necessary step and often considered as one of the bottlenecks of the expansion of visual servoing. That is why more and more non geometric visual measurements are proposed [3].

3.5. Sparse Representations and Bayesian model selection

Sparse representation methods aim at finding representations of a signal with a small number of components taken from an over-complete dictionary of elementary functions or vectors. Sparse representation are of interest in a number of applications in Physics and signal processing. In particular, they provide a simple characterization of certain families of signals encountered in practice. For example, smooth signals can be shown to have a sparse representation in over-complete Fourier or wavelet dictionaries. More recently, it has been emphasized in [42] that the solutions of certain differential equations (*e.g.,* diffusion or transport equation) have a sparse representation in dictionaries made up of curvelets.

Finding the sparse representation of a signal typically requires to solve an under-determined system of equations under the constraint that the solution is composed of the minimum number of non-zero elements. Unfortunately, this problem is known to be NP-hard and sub-optimal procedures have to be devised to find practical solutions. Among the various algorithms that find approximate solutions, let us mention for example the matching pursuit, orthogonal matching pursuit or basis pursuit algorithms.

Choosing appropriate models and fixing hyper-parameters is a tricky and often hidden process in optic-flow estimation. Most of the motion estimators proposed so far have generally to rely on successive trials and a empirical strategy for fixing the hyper parameters values and choosing the adequate model. Besides of its computational inefficiency, this strategy may produce catastrophic estimate without any relevant feedback for the end-user, especially when motions are difficult to apprehend as for instance for complex deformations or non-conventional imagery. Imposing hard values to these parameters may also yield poor results when the lighting conditions or the underlying motions differ from those the system has been calibrated with. At the extreme, the estimate may be either too smooth or at the opposite non-existent strong motion discontinuities.

Bayesian model selection offers an attractive solution to this problem. The Bayesian paradigm implicitly requires the definition of several competing observation and prior *probabilistic* model(s). The observation model relates the motion of the physical system to the spatial and temporal variations of the image intensity. The prior models define the spatio-temporal constraints that the motion have to satisfy. Considering these competing models, the Bayesian theory provides methodologies to select the best models under objective performance criterion (minimum probability of error, minimum mean square error, etc). Moreover, due to the generality of this problem, numerous algorithms and approximations exist in the literature to implement efficient and effective practical solutions: Monte-Carlo integration, mean-field and Laplace approximations, EM algorithm, graphical models, etc.

GALEN Team

3. Scientific Foundations

3.1. Structured coupled low- and high-level visual perception

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:

$$Y^* = \text{argmax}_Y s(X, Y) \tag{90}$$

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 .

A driving force behind research in GALEN has been the understanding that these three aspects are tightly coupled. In particular, efficient optimization can be achieved by resorting to sparse image representations that 'shortlist' putative solutions and/or by working with scoring functions that can be efficiently optimized. However, the accuracy of a scoring function is largely affected by the breadth of relationships that it accommodates, as well as the completeness of the employed image representation. Determining the tradeoff between these two requirements is far from obvious and often requires approaches customized to the particular problem setting addressed. Summarizing, even though the three problems outlined above can be addressed in isolation, an integrated end-to-end approach is clearly preferable, both for computational efficiency and performance considerations.

Research in GALEN has therefore dealt with the following problem aspects: first, developing a generic and reliable low-level image representation that can be used transversally across multiple tasks. The use of learning-based techniques has been pursued for boundary detection and symmetry detection in [32], yielding state-of-the-art results, while in [27] trajectory grouping was used to come up with a mid-level representation of spatio-temporal data. Complementary to the detection of geometric structures, we have also explored methods for their description both for image and surface data [17]. We are currently pursuing the formulation of the task in structured prediction terms, which will hopefully allow us to exploit the geometrical interdependencies among symmetry and boundary responses.

Second, we have worked on learning scoring functions for detection with deformable models that can leverage upon the developed low-level representations, while also being amenable to efficient optimization. Building on our earlier work on using boundary and symmetry detector responses to perform groupwise registration within categories we used discriminative learning to train hierarchical object models that rely on shape-based representations; these were successfully applied to the detection of shape-based categories, while we are currently pursuing their integration with appearance-based models.

Third, efficient optimization for deformable models was pursued in [18], where we have developed novel techniques for object detection that employ combinatorial optimization tools (A* and Branch-and-Bound) to tame the combinatorial complexity; in particular our work has a best-case performance that is logarithmic in the number of pixels, while our work in [18] allows us to further accelerate object detection by integrating low-level processing (convolutions) with a bounding-based object detection algorithm. Working on a different approach, in [10] we have pursued the exploitation of reinforcement-learning to optimize over the set of shapes derivable from shape grammars. We are currently pursuing a full-fledged bounding-based inference algorithm, which will integrate the tasks of boundary detection and grouping in a single, integrated object detection algorithm.

3.2. Machine Learning & Structure 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

$$\mathcal{R}(f) = \int \ell(f, x, y) dP(x, y), \tag{91}$$

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,

$$\mathcal{R}(f) \approx \widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, x_i, y_i), \tag{92}$$

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:

$$\min_{f \in \mathcal{F}} \lambda \Omega(\|f\|) + \widehat{\mathcal{R}}(f), \tag{93}$$

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:

$$\min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}) + \sum_{i=1}^{n} \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \tag{94}$$

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:

$$\min_{\mathbf{w}\in\mathcal{W},\mathbf{v}\in\{0,1\}^n} R(\mathbf{w}) + \sum_{i=1}^{n} v_i \Delta(y_i, y_i(\mathbf{w}), h_i(\mathbf{w})) - \mu_t \sum_{i=1}^{n} v_i. \tag{95}$$

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,

$$\min_{\mathbf{w}\in\mathcal{W},\theta\in\Theta} R(\mathbf{w}) + \sum_{i=1}^{n} \sum_{h_i\in\mathcal{H}} \Pr(h_i|x_i, y_i; \theta) \Delta(y_i, h_i, y_i(\mathbf{w}), h_i(\mathbf{w})). \tag{96}$$

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:

$$\arg\min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}). \tag{97}$$

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:

$$\arg\min_{\{x_p\}} \mathcal{P}(g, f) = \sum_{p \in \mathcal{V}} g_p(x_p) + \sum_{(p_1, p_2) \in \mathcal{E}} f_{p_1 p_2}(x_{p_1}, x_{p_2}) + \sum_{C_i \in \mathcal{E}} f_{p_1 \cdots p_n}(x_{p_{i^1}}, \cdots, p_{x_{i|C_i|}}). \tag{98}$$

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 Team

3. Scientific Foundations

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 fingerprints. 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. Scientific Foundations

3.1. Modeling of bacterial regulatory networks

Participants: Sara Berthoumieux, Eugenio Cinquemani, Johannes Geiselmann, Nils Giordano, Edith Grac, Hidde de Jong, Stéphane Pinhal, Delphine Ropers [Correspondent], Valentin Zulkower.

The adaptation of bacteria to changes in their environment is controlled on the molecular level by large and complex interaction networks involving genes, mRNAs, proteins, and metabolites (Figure 2). The elucidation of the structure of these networks has much progressed as a result of decades of work in genetics, biochemistry, and molecular biology. Most of the time, however, it is not well understood how the response of a bacterium to a particular environmental stress emerges from the interactions between the molecular components of the network. This has called forth an increasing interest in the mathematical modeling of the dynamics of biological networks, in the context of a broader movement called systems biology.

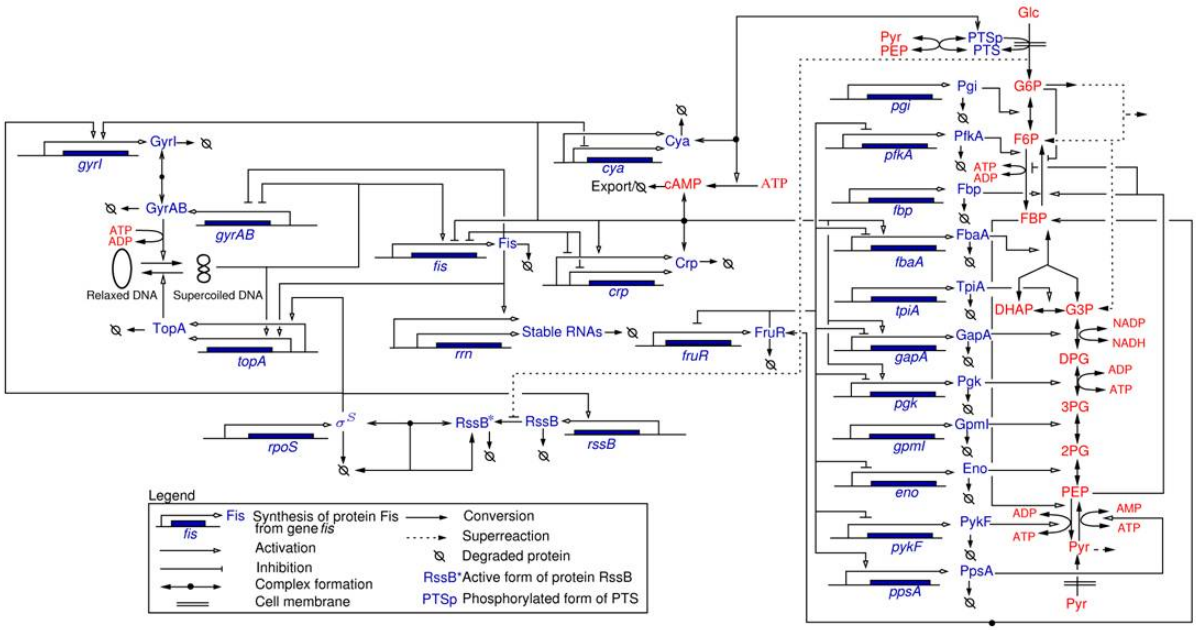


Figure 2. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in E. coli (Baldazzi et al., PLoS Computational Biology, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description of the 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.

In theory, it is possible to write down mathematical models of biochemical networks, and study these by means of classical analysis and simulation tools. In practice, this is not easy to achieve though, as quantitative data on kinetic parameters are usually absent for most systems of biological interest. Moreover, the models include a large number of variables, are strongly nonlinear and include different time-scales, which make them difficult to handle both mathematically and computationally. A possible approach to this problem has been to use approximate models that preserve essential dynamical properties of the networks. Different approaches have been proposed in the literature, such as the use of approximations of the typical response functions found in gene and metabolic regulation and the reduction of the model dimension by decomposing the system into fast and slow subsystems. These reductions and approximations result in simplified models that are easier to analyze mathematically and for which parameter values can be more reliably estimated from the available experimental data.

Several modeling approaches are exploited in IBIS to gain a better understanding of the ability of *E. coli* to adapt to a various nutritional and other environmental stresses, such as carbon, phosphate, and nitrogen starvation. We are particularly interested in the role of networks of global regulators in shaping the adaptive response of bacteria. Moreover, we study the interactions of these networks with metabolism and the gene expression machinery. These topics involve collaborations with the BEAGLE, COMORE, and CONTRAINTES project-teams at Inria.

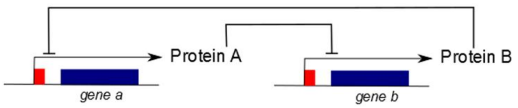
3.2. Analysis, simulation, and identification of bacterial regulatory networks

Participants: Sara Berthoumieux, Eugenio Cinquemani, Johannes Geiselmann, Nils Giordano, Hidde de Jong [Correspondent], Michel Page, François Rechenmann, Delphine Ropers, Diana Stefan, Valentin Zulkower.

Computer simulation is a powerful tool for explaining the capability of bacteria to adapt to sudden changes in their environment in terms of structural features of the underlying regulatory network, such as interlocked positive and negative feedback loops. Moreover, computer simulation allows the prediction of unexpected or otherwise interesting phenomena that call for experimental verification. The use of simplified models of the stress response networks makes simulation easier in two respects. In the first place, model reduction restricts the class of models to a form that is usually easier to treat mathematically, in particular when quantitative information on the model parameters is absent or unreliable. Second, in situations where quantitative precision is necessary, the estimation of parameter values from available experimental data is easier to achieve when using models with a reduced number of parameters.

Over the past few years, we have developed in collaboration with the COMORE project-team a qualitative simulation method adapted to a class of piecewise-linear (PL) differential equation models of gene regulatory networks. The PL models, originally introduced by Leon Glass and Stuart Kauffman, provide a coarse-grained picture of the dynamics of gene regulatory networks. They associate a protein or mRNA concentration variable to each of the genes in the network, and capture the switch-like character of gene regulation by means of step functions that change their value at a threshold concentration of the proteins. The advantage of using PL models is that the qualitative dynamics of the high-dimensional systems are relatively simple to analyze, using inequality constraints on the parameters rather than exact numerical values. The qualitative dynamics of gene regulatory networks can be conveniently analyzed by means of discrete abstractions that transform the PL model into so-called state transition graphs.

The development and analysis of PL models of gene regulatory network has been implemented in the qualitative simulation tool GENETIC NETWORK ANALYZER (GNA) (Section 4.1). GNA has been used for the analysis of several bacterial regulatory networks, such as the initiation of sporulation in *B. subtilis*, quorum sensing in *P. aeruginosa*, the onset of virulence in *E. chrysanthemi*, and environmental biodegradation by *P. putida* mt-2. GNA is currently distributed by the Genostar company, but remains freely available for academic research. The analysis of models of actual bacterial regulatory networks by means of GNA leads to large state transition graphs, which makes manual verification of properties of interest practically infeasible. This has motivated the coupling of GNA to formal verification tools, in particular model checkers that allow properties formulated in temporal logic to be verified on state transition graphs. This has been the subject of collaborations with the POP-ART and VASY project-teams at Inria Grenoble - Rhône-Alpes.



(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 3. (a) 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. (b) 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 ($\gamma_a \, x_a$).

Recent advances in experimental techniques have led to approaches for measuring cellular processes in real-time on the molecular level, both in single cells and populations of bacteria (Section 3.3). The data sources that are becoming available by means of these techniques contain a wealth of information for the quantification of the interactions in the regulatory networks in the cell. This has stimulated a broadening of the methodological scope of IBIS, from qualitative to quantitative models, and from PL models to nonlinear ODE models and even stochastic models. The group has notably started to work on what is the bottleneck in the practical use of these models, the structural and parametric identification of bacterial regulatory networks from time-series data, in collaboration colleagues from INRA, the University of Pavia (Italy) and ETH Zürich (Switzerland). This raises difficult problems related to identifiability, measurement noise, heterogeneity of data sources, and the design of informative experiments that are becoming increasingly prominent in the systems biology literature.

3.3. High-precision measurements of gene expression in bacteria

Participants: Guillaume Baptist, Sara Berthoumieux, Julien Demol, Johannes Geiselmann [Correspondent], Edith Grac, Jérôme Izard, Hidde de Jong, Stephan Lacour, Yves Markowicz, Corinne Pinel, Stéphane Pinhal, Delphine Ropers, Claire Villiers, Valentin Zulkower.

The aim of a model is to describe the functioning of bacterial regulatory networks so as to gain a better understanding of the molecular mechanisms that control cellular responses and to predict the behavior of the system in new situations. In order to achieve these goals, we have to calibrate the model so that it reproduces available experimental data and confront model predictions with the results of new experiments. This presupposes the availability of high-precision measurements of gene expression and other key processes in the cell.

We have notably resorted to the measurement of fluorescent and luminescent reporter genes, which allow monitoring the expression of a few dozens of regulators in parallel, with the precision and temporal resolution needed for the validation of our models. More specifically, we have constructed transcriptional and translational fusions of key regulatory genes of *E. coli* to fluorescent and luminescent reporter genes (Figure 4). The signals of these reporter genes are measured *in vivo* by an automated, thermostated microplate reader. This makes it possible to monitor in real time the variation in the expression of a few dozens of genes in response to an external perturbation. We have developed an experimental pipeline that resolves most technical difficulties in the generation of reproducible time-series measurements. The pipeline comes with data analysis software that converts the measurements into representations of the time-course of promoter activities that

can be compared with model predictions (Section 4.2). In order to obtain rich information about the network dynamics, we have begun to measure the expression dynamics in both wild-type and mutant cells, using an existing *E. coli* mutant collection. Moreover, we have developed tools for the perturbation of the system, such as expression vectors for the controlled induction of particular genes.

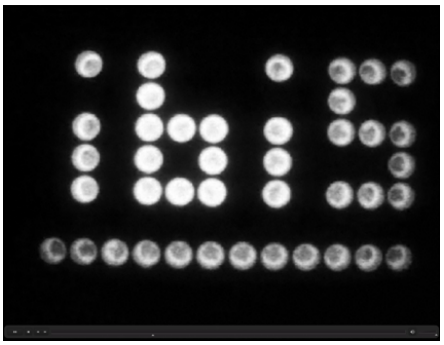


Figure 4. Playful illustration of the principle of reporter genes (see http://ibis.inrialpes.fr for the corresponding movie). A microplate containing a minimal medium (with glucose and acetate) is filmed during 36 hours. Wells contain E. coli bacteria which are transformed with a reporter plasmid containing the luciferase operon (luxCDABE) under control of the acs promoter. This promoter is positively regulated by the CRP-cAMP complex. When bacteria have metabolized all the glucose, the cAMP concentration increases quickly and activates the global regulator CRP which turns on the transcription of the luciferase operon producing the light. The glucose concentration increases from left to right on the microplate, so its consumption takes more time when going up the gradient and the letters appear one after the other. The luciferase protein needs reductive power (FMNH2) to produce light. At the end, when acetate has been depleted, there is no more carbon source in the wells. As a consequence, the reductive power falls and the "bacterial billboard" switches off. Source: Guillaume Baptist.

While reporter gene systems allow the dynamics of gene expression to be measured with high precision and temporal resolution on the level of cell populations, they do not provide information on all variables of interest though. Additional technologies may complement those that we have developed in our laboratory, such as the tools from transcriptomics, proteomics, and metabolomics that are able to quantify the amounts of mRNAs, proteins and metabolites, respectively, in the cells at a given time-point. In addition, for many purposes it is also important to be able to characterize gene expression on the level of single cells instead of cell populations. This requires experimental platforms that measure the expression of reporter genes in isolated cells by means of fluorescence and luminescence microscopy. IBIS has access to these technologies through collaborations with other groups on the local and national level, such as the INSA de Toulouse and the Laboratoire Interdisciplinaire de Physique at the Université Joseph Fourier.

MACS Project-Team

3. Scientific Foundations

3.1. Formulation and analysis of effective and reliable shell elements

Thin structures (beams, plates, shells...) are widely considered in engineering applications. However, most experts agree that the corresponding discretization procedures (finite elements) are not yet sufficiently reliable, in particular as regards shell structures. A major cause of these difficulties lies in the numerical locking phenomena that arise in such formulations [2].

The expertise of the team in this area is internationally well-recognized, both in the mathematical and engineering communities. In particular, we have strongly contributed in analysing – and better explaining – the complex locking phenomena that arise in shell formulations [2]. In addition, we have proposed the first (and only to date) shell finite element procedure that circumvents locking [6]. However, the specific treatment applied to avoid locking in this procedure make it unable to correctly represent membrane-dominated behaviors of structures (namely, when locking is not to be expected). In fact, a "perfect shell element" – namely, with the desired reliability properties mathematically substantiated in a general framework – is still to be discovered, whereas numerous teams work on this issue throughout the world.

Another important (and related) issue that is considered in the team pertains to the design and analysis of numerical procedures that are adapted to industrial applications, i.e. that fulfill some actual industrial specifications. In particular, in the past we have achieved the first mathematical analysis of "general shell elements" – which are based on 3D variational formulations instead of shell models – these elements being among the most widely used and most effective shell elements in engineering practice.

3.2. Stability and control of structures

Stability of structures is – of course – a major concern for designers, in particular to ensure that a structure will not undergo poorly damped (or even unbounded) vibrations. In order to obtain improved stability properties – or to reach nominal specifications with a thinner a lighter design – a control device (whether active, semi-active, or passive) may be used.

The research performed in the team in this area – other than some prospective work on robust control – has been so far primarily focused on the stability of structures interacting with fluid flows. This problem has important applications e.g. in aeronautics (flutter of airplane wings), in civil engineering where the design of long-span bridges is now partly governed by wind effects, and in biomechanics (blood flows in arteries, for instance). Very roughly, the coupling between the structure and the flow can be described as follows: the structural displacements modify the geometry of the fluid domain, hence the fluid flow itself which in turn exerts an action on the structure. The effects of structural displacements on the fluid can be taken into account using ALE techniques, but the corresponding direct simulations are highly CPU-intensive, which makes stability analyses of such coupled problems very costly from a computational point of view. In this context a major objective of our work has been to formulate a simplified model of the fluid-structure interaction problem in order to allow computational assessments of stability at a reasonable cost.

3.3. Modeling and estimation in biomechanics

A keen interest in questions arising from the need to model biomechanical systems – and to discretize such problems – has always been present in the team since its creation. Our work in this field until now has been more specifically focused on the objectives related to our participation in the ICEMA ARC projects and in the CardioSense3D initiative, namely, to formulate a complete continuum mechanics model of a beating heart, and to confront – or "couple", in the terminology of the Inria strategic plan – numerical simulations of the model with actual clinical data via a data assimilation procedure.

Our global approach in this framework thus aims at using measurements of the cardiac activity in order to identify the parameters and state of a global electromechanical heart model, hence to give access to quantities of interest for diagnosing electrical activation and mechanical contraction symptoms. The model we propose is based on a chemically-controlled constitutive law of cardiac myofibre mechanics consistent with the behavior of myosin molecular motors [27]. The resulting sarcomere dynamics is in agreement with the "sliding filament hypothesis" introduced by Huxley. This constitutive law has an electrical quantity as an input which can be independently modeled, considered as given (or measured) data, or as a parameter to be estimated.

MAGIQUE-3D Project-Team

3. Scientific Foundations

3.1. Inverse Problems

- **Inverse scattering problems.** The determination of the shape of an obstacle from its effects on known acoustic or electromagnetic waves is an important problem in many technologies such as sonar, radar, geophysical exploration, medical imaging and nondestructive testing. This inverse obstacle problem (IOP) is difficult to solve, especially from a numerical viewpoint, because it is ill-posed and nonlinear [77]. Moreover the precision in the reconstruction of the shape of an obstacle strongly depends on the quality of the given far-field pattern (FFP) measurements: the range of the measurements set and the level of noise in the data. Indeed, the numerical experiments (for example [88], [91], [84], [85]) performed in the resonance region, that is, for a wavelength that is approximately equal to the diameter of the obstacle, tend to indicate that in practice, and at least for simple shapes, a unique and reasonably good solution of the IOP can be often computed using only one incident wave and *full aperture* far-field data (FFP measured only at a limited range of angles), as long as the aperture is larger than π. For smaller apertures the reconstruction of the shape of an obstacle becomes more difficult and nearly impossible for apertures smaller than $\pi/4$.
  This plus the fact that from a mathematical viewpoint the FFP can be determined on the entire sphere $S1$ from its knowledge on a subset of $S1$ because it is an *analytic* function, we propose [74], [75] a solution methodology to extend the range of FFP data when measured in a limited aperture and not on the entire sphere $S1$. It is therefore possible to solve the IOP numerically when only limited aperture measurements are available. The objective of MAGIQUE-3D is to extend this work to 3D problems of acoustic scatterin and to tackle the problem of elasto-acoustic scattering.
- **Depth Imaging in the context of DIP.** The challenge of seismic imaging is to obtain the best 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 most used technique of imaging is RTM (Reverse Time Migration), [76], which 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. The research program DIP has been defined by researchers of MAGIQUE-3D and engineers of TOTAL jointly. It has been created with the aim of gathering researchers of Inria, with different backgrounds and the scientific programm will be coordinated by MAGIQUE-3D. In this context, MAGIQUE-3D will contribute by working on the inverse problem and by continuing to develop 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 focus 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 Schemes in Space and Time.** 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 considerably reduce the space step, which is the distance between two points of the mesh representing the computational domain. Obviously this results in increasing the number of unknowns of the discrete problem. Besides, the time step, whose value fixes the number of required iterations for solving the evolution problem, is linked to the space step through

the 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 (and of multiplications by the stiffness matrix) will be. The method that we proposed in [11] allows for the use of local time-step, adapted to the various sizes of the cells and we recently extended it to deal with p-adaptivity [73]. However, this method can not yet handle dissipation terms, which prevents us for using absorbing boundary conditions or Perfectly Matched Layers (PML). To overcome this difficulty, we will first tackle the problem to used the modified equation technique [78], [90], [82] with dissipation terms, which is still an open problem.

We are also considering an alternative approach to obtain high-order schemes. The main idea is to apply first the time discretization thanks to the modified equation technique and after to consider the space discretization. Our approach involves p-harmonic operators, which can not be discretized by classical finite elements. For the discretization of the biharmonic operator in an homogeneous acoustic medium, both C1 finite elements (such as the Hermite ones) and Discontinuous Galerkin Finite Elements (DGFE) can be used while in a discontinuous medium, or for higher-order operators, DGFE should be preferred [68]. This new method seems to be well-adapted to p-adaptivity. Therefore, we now want to couple it to our local time-stepping method in order to deal with hp-adaptivity both in space and in time. We will then carry out theoretical and numerical comparisons between this technique and the classical modified equation scheme.

Once we have performant hp-adaptive techniques, it will be necessary to obtain error-estimators. Since we consider huge domain and complex topography, the remeshing of the domain at each time-step is impossible. One solution would be to remesh the domain for instance each 100 time steps, but this could also hamper the efficiency of the computation. Another idea is to consider only p adaptivity, since in this case there is no need to remesh the domain.

- **Mixed hybrid finite element methods for the wave equation.** The new mixed-hybrid-like method for the solution of Helmholtz problems at high frequency we have built enjoys the three following important properties: (1) unlike classical mixed and hybrid methods, the method we proposed is not subjected to an inf-sup condition. Therefore, it does not involve numerical instabilities like the ones that have been observed for the DGM method proposed by Farhat and his collaborators [80], [81]. We can thus consider a larger class of discretization spaces both for the primal and the dual variables. Hence we can use unstructured meshes, which is not possible with DGM method (2) the method requires one to solve Helmholtz problems which are set inside the elements of the mesh and are solved in parallel(3) the method requires to solve a system whose unknowns are Lagrange multipliers defined at the interfaces of the elements of the mesh and, unlike a DGM, the system is hermitian and positive definite. Hence we can use existing numerical methods such as the gradient conjugate method. We intend to continue to work on this subject and our objectives can be described following three tasks: (1) Follow the numerical comparison of performances of the new methods with the ones of DGM. We aim at considering high order elements such as R16-4, R32-8, ...; (2)Evaluate the performance of the method in case of unstructured meshes. This analysis is very important from a practical point of view but also because it has been observed that the DGM deteriorates significantly when using unstructured meshes; (3) Extend the method to the 3D case. This is the ultimate objective of this work since we will then be able to consider applications.

  Obviously the study we propose will contain a mathematical analysis of the method we propose. The analysis will be done in the same time and we aim at establishing a priori and a posteriori estimates, the last being very important in order to adopt a solution strategy based on adaptative meshes.

- **Boundary conditions.** The construction of efficient absorbing conditions is very important for solving wave equations, which are generally set in unbounded or very large domains. The efficiency of the conditions depends on the type of waves which are absorbed. Classical conditions absorb propagating waves but recently new conditions have been derived for both propagating and evanescent waves in the case of flat boundaries. MAGIQUE-3D would like to develop new absorbing boundary conditions whose derivation is based on the full factorization of the wave equation using pseudodifferential calculus. By this way, we can take the complete propagation phenomenon into account

which means that the boundary condition takes propagating, grazing and evanescent waves into account, and then the absorption is optimized. Moreover our approach can be applied to arbitrarily-shaped regular surfaces.
We intend to work on the development of interface conditions that can be used to model rough interfaces. One approach, already applied in electromagnetism [89], consists in using homogenization methods which describes the rough surface by an equivalent transmission condition. We propose to apply it to the case of elastodynamic equations written as a first-order system. In particular, it would be very interesting to investigate if the rigorous techniques that have been used in [70], [71] can be applied to the theory of elasticity. This type of investigations could be a way for MAGIQUE-3D to consider medical applications where rough interfaces are often involved. Indeed, we would like to work on the modeling and the numerical simulation of ultrasonic propagation and its interaction with partially contacting interfaces, for instance bone/titanium in the context of an application to dentures, in collaboration with G. Haiat (University of Paris 7).

- **Asymptotic modeling**.

  In the context of wave propagation problems, we are investigating physical problems which involves multiple scales. Due to the presence of boundary layers (and/or thin layers, rough interfaces, geometric singularities), the direct numerical simulation (DNS) of these phenomenas involves a large numbers of degrees of freedom and high performance computing is required. The aim of this work is to develop credible alternatives to the DNS approach. Performing a multi-scale asymptotic analysis, we derive approximate models whose solution can be computed for a low computational cost. We study these approximate models mathematically (well-posedness, uniform error estimates) and numerically (we compare the solution of these approximate models to the solution of the initial model computed with high performance computating).

  We are mostly interested in the following problems.

  – 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;

  – ultrasonic wave propagation through bone-titanium media in medicine in collaboration with Inria Team MC2, and MSME;

  – asymptotic modeling of multi perforate plates in turbo reactors in collaboration with Cerfacs, INSA-Toulouse, Onera and Snecma in the framework of the ANR APAM.

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 calculus can be performed locally on each element of the mesh. The communications are ensured 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 to 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 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.** We are now planning to port the code on the new Intel Many Integrated Core Architecture (Intel MIC). The optimization of this code should begin in 2013, in collaboration with Dider Rémy of SGI.

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 impossible to reach in the past.

MAGNOME Project-Team

3. Scientific Foundations

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 [20], [6], [10], [11], [56], [57].
- **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], [4], [41], [44], [29].
- **Ancestor reconstruction** using optimization techniques, to provide plausible scenarios of the history of genome evolution [11], [8], [46], [62].
- **Classification and logical inference**, to reliably identify similarities between groups of genetic elements, and infer rules through deduction and induction [9], [7], [10].
- **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 [40], [27], [60], [34], [59].

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 [56], [57], [6], [10], [2], [11] and to
- prokaryotes from the lactic bacteria used in winemaking [20], [23], [21], [32], [36], [28].

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 BioRica [26], [40], [60], a framework in which discrete and continuous models can communicate with a clear semantics. Hierarchical models in BioRica can be assembled from existing models, and translated into their execution semantics and then simulated at multiple resolutions through multi-scale stochastic simulation. BioRica 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 [5] [27] and inferring new models through a combination of comparative genomics and successive refinement [51], [52].

MASAIE Project-Team

3. Scientific Foundations

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, allow 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 [33]. 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 [27], [29]. 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) will be 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 is 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

$$\begin{cases} \dot{x} = f(x,u) \\ y = h(x), \end{cases} \tag{99}$$

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 the system (1) is a dynamical system

$$\dot{\widehat{x}}(t) = g(\widehat{x}(t), y(t), u(t)), \tag{100}$$

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 $\widehat{x}(0)$

$$\|x(t) - \widehat{x}(t)\| \leq c\, \|x(0) - \widehat{x}(0)\|\, e^{-a\,t}\ ,\ \ \forall t > 0.$$

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. Gouze, O. Bernard, G. Sallet) and is still very active in this area. Now the problem is to identify relevant class 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 there is only one case of use of an observer in virology (Velasco-Hernandez J. , Garcia J. and Kirschner D. [38]) in modeling the chemotherapy of HIV, but this observer, based on classical linear theory, 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 becomes a near certainty. The rationale for studying continuous delays are 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 approaches (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 [5].

MNEMOSYNE Team

3. Scientific Foundations

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 [31]. 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 [27], 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 [33], [21]. 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 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 developped for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [18], [26] and were the basis for several contributions in our group [32], [29], [34]. 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 [19]; the associated formalism [24] 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 developped for dynamic neural fields [16], that demonstrated a richness of behavior [20] 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 [29], [30] and their use for observing the emergence of typical cortical properties [23]. 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 [25] 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 for 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 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 [22], [15]. 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 skills that have been learned previously. 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 also 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 [17], 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. Scientific Foundations

3.1. Modelling and simulating microbial ecosystems

Microbial ecosystems naturally put into play phenomena at different scales, from the individual level at a microscopic scale to the population level at a macroscopic scale, with sometimes intermediate levels. The size of substrate molecules is a thousand time smaller than the size of microorganisms and usually diffuse much faster. The substrate consumption of one microorganism is negligible at the population level but the sum of the consumption of its neighbors can modify the local concentration of substrate, which itself modifies microorganism growth, acting as a *feedback loop*. For other variables that change slowly (pH, temperature...) cumulative effects create intermediate time scales, coupling individual and environment dynamics. The very large populations justify macroscopic modelling but for some ecosystems, spatial structures seen at intermediate scale need to be tackled. This is typically the case of biofilm ecosystems, for which the biofilm structure is responsible of characteristics of the overall ecosystem. Models that are purely individual-based or purely populational are rarely truly satisfactory to incorporate current knowledge on microbial ecosystems at various scales and to push ahead mathematical analysis or to derive operational rules.

3.1.1. Macroscopic models

The starting point is the knowledge of biologists that report a large number of mechanisms discovered or shown on laboratory experiments at a population level, such as competition for a growth-limiting substrate, predation interactions, obligate mutualism or communication between bacteria. If each *elementary* mechanism is today well understood and modelled at a macroscopic level, the consideration of several mechanisms together in a single model is still raising several questions of understanding and prediction. This is typically the case when there is more than one growth-limiting substrate in the chemostat model or when one couples species competition with a spatial structure (flocculation, niches...).

1. *Non-spatial models.*
   Ordinary differential equations (ODE) are the common way to describe the evolution of the size or concentration of species populations and their functional contribution in resource transformation (such as substrate degradation) in homogeneous or perfectly mixed compartments (or ecological niches). The well-known chemostat model [87] used in microbiology for single strain:

$$\begin{aligned} \dot{s} &= -\tfrac{1}{y}\mu(s)\,b + D\,(s_{in} - s) \\ \dot{b} &= \mu(s)\,b - D\,b \end{aligned}$$

   (where s and b stand respectively for the substrate and biomass concentrations, $\mu(s)$ is specific growth function, D the dilution rate, s_{in} the input substrate concentration), has to be extended to cope with the specificity of microbial ecosystems in the following directions.

   - very large number (hundreds or thousands) of species. This leads to the problem of characterization of their distribution during the transients, that is a way to study the *functional redundancy* of ecosystems.
   - environmental fluctuations (input flow rate, input concentration, temperature, pH...). This impacts the efficiency of a microbial ecosystem, when biological and environmental time scales are different. *Singular perturbations* is the technique we use to separate *slow* variables from *fast* ones, leading to approximations of the dynamics on *slow manifolds* to be determined and analyzed.

- interactions due to several limited resources and trophic chains. Most of the literature on the chemostat considers models with single limited resource, while some work studied purely essential or substitutable resources.
- several populations of bacteria (for each species) to describe the effects of certain spatial structures that are artificially created in bioreactors or naturally found in soils, like flocks, colonies or biofilms: the planktonic (or free) cells and the biofilm (or fixed) biomass (for telluric ecosystems, such a distinction is also relevant to represent the sticking/non sticking characteristics of soil). Considering simple models of aggregates (that are not spatialized) can provide a simplified model of the dynamics of the overall biomass.
- *active* and *dormant* bacteria. This distinction is motivated by the observations made on ecosystems of sparse resources such as arid soils.

2. *Spatial models.*
   In the spirit of lattice differential equations, representations in terms of networks of (abstract) interconnected bioreactors propose an intermediate level between models of average biomass (a single ODE) and a continuous representation of space (PDE). A model of interconnected bioreactors is a way to *implicitly* take into account spatial heterogeneity, without requiring a precise knowledge of it. It is similar to the island models used in ecology [85] but coupled with the dynamics of abiotic resources and hydrodynamics laws (transport, percolation, diffusion) governing the transfers between patches. This approach appears to be relevant for telluric ecosystems, for which pedologists report that microbial activities in soil are usually concentrated in *hot-spots* that could be seen as small bioreactors. Understanding the role of the topology of the interconnection network and how a spatial structure impacts the outputs is also relevant in biotechnology to improve the yield or stability of processes.

3.1.2. Microscopic models

In these models (birth and death processes, neutral models, individual-based models) the dynamic of the population is described in terms of discrete events: birth and death of individuals, or jumps in terms of biomass. These models can be gathered under the same framework that could be called *Markov stochastic processes with discrete events*. Most of the time they should be coupled with continuous components like the size of each individual or the dynamic of the resources (represented in terms of ODE or PDE).

The Markovian framework allows on the one hand sharp analyses and rescaling techniques [76]; on the other hand it induces a simplification in the memory structure of the process that is important in terms of simulation. Indeed, as the future state of the system depends from the past only through the present state, only the current state should be kept in memory for simulation.

We will consider three families of processes with discrete events, from simplest to most complex.

1. *Birth and death processes.*
   These models [77] are of first importance in small population size. They indeed allow investigation of near-to-extinction situations in a more realistic ways than the classical ODE models: they permit the computation, analytically but most of the time numerically, the distribution of extinction time and the probability of extinction. Efforts should be made to developed efficient Monte Carlo simulation procedures and approximation techniques for extinction probability and time distribution evaluation. In larger population sizes, they are advantageously approximated by diffusion models (see next section).
2. *The neutral models.*
   In *neutral* models [82] sizes of different species evolve as birth and death processes with immigration: all individuals have the same characteristics and are not spatialized. Such hypotheses could be considered unrealistic from a purely biological perspective, but these models focus on some precise properties to be simulated and predicted (for instance the biodiversity).

Comparing the prediction of species abundance of these models to real observations provides a way to justify or invalidate the neutral hypothesis. Extensions of the neutral model, that was originally introduced for forest ecology, have to be developed in order to better suit the framework of microbial ecology, such as the non constant size of the populations and spatialized variations.

3. *The individual based models.*
   IBM's [68], [69] appear to be well suited to describe colonies or biofilms [88]: in addition to birth, death and movement events, one has to consider *aggregation* and *detachment* events. The mechanisms that lead to the emergence of spatial patterns of colonies, or the formation of biofilms, which adhere to surface via polymers generated by the bacteria under specific hydrodynamics conditions, are not well understood yet. Typically, one can consider that bacteria inside the aggregates are disadvantaged to access the nutrient.

IBM modelling is a convenient way to propose aggregation and detachment mechanisms at the individual level in terms of random events connected to the geometry of the neighborhood, and to compare generated images with microscopic observations (for instance the confocal microscopy).

One has to be aware that few methods are available to study systematically and rigorously the properties of IBM, contrary to models based on differential equations (ODE, PDE...).

3.1.3. Bridges between models

The "theory of a computational model", that combine two kinds of models (typically ODE and IBM) that are different representations of the same objects, relies on two steps: the "program making" and the "theoretical study", in the spirit of the *double modelling* approach (roughly speaking, it consists in grasping the complexity of a IBM by analyzing accurately the consequences of each hypothesis on the macroscopic behavior of the model, building an approximate model of its global dynamics). Two main tools can be considered.

1. *Change of scale.*
   For IBM models (neutral or Markovian), we consider mean field and moments approximation techniques that provide information at the macroscopic (i.e. populational) level, to be compared with macroscopic models. From a birth-and-death process describing the individual level, a renormalisation can provide a stochastic differential equation at a meso-scale. The *diffusion approximation* technique can be understood as a numerical acceleration technique where the number of births and deaths follows a normal law. These stochastic models at meso-scale can provide additional information compared to deterministic models at a macro-scale, such as parameter identifiability or finite time extinction. The price to pay is to give much more conceptual and numerical efforts, that become less relevant for very large populations.

   For PDE models on spatial domains described with regular patterns (such as models of biofilm), the homogenization technique allows to obtain simpler PDE with constant parameters.

2. *The multi-scale modelling.*
   The spatial heterogeneity in microbial ecosystems requires to consider simultaneously several scales:
   - a *physical* scale. In batch processes, nutrient diffusion can be modelled by adapting the heat equation with Dirichlet boundary conditions. In continuous reactors, a convection-diffusion equation with Neumann boundary conditions is considered instead, the speed vector field being provided by the equations of fluid mechanics. The spatial scale used for the discretization is given by diffusion and flow parameters.
   - a *biological* scale, given by the size and mobility of bacteria. Usually, this scale is larger than the physical one (at least in the liquid phase).
   - an *aggregation* scale of colonies or biofilms, even larger, that provides the spatial patterns.

It is always possible to describe all the processes at the smaller common scale and then use a global representation, but this leads to extremely long computation times. The challenge is to manage these overlapping scales together and guarantee the stability of the numerical schemes. This is the goal of the *multi-scale* approaches [84]. For microbial ecosystems, it consists in

1. proposing new representations of the various scales of aggregation of bacteria in a model, taking into account the attachment-detachment processes determined by the local hydrodynamics conditions. Here, discussions with specialists of fluid mechanics are required.
2. coupling diversity models (e.g. models based on the neutral assumption) with spatial models (that reproduce the patterns observed on images of microscopy) to better understand the link biodiversity/structure.
3. introducing new *control* variables, considered as independent variables, each of them describing a proper scale. For this purpose, we investigate different techniques available to determine such variables: *mean-field* approximation, *singular perturbations*, *unification by limiting layers* or
4. *renormalising*, that aims at detecting invariants among models of different scales.

3.2. Interpreting and analyzing experimental observations

The validation of microbial models on data is rarely a straightforward task, because observations are most of the time not directly related to the variables of the models. Techniques such as abundance spectrum provided by molecular biology or confocal imagery are relatively recent in the field of microbial ecosystems. The signals provided by theses devices leave many research questions open in terms of data interpretation and experiments design. One can distinguish three kinds of key information that are needed at the basis of model assumptions:

- structure of the communities (i.e. who is present?),
- nature of interactions between species (competition, mutualism, syntrophism...),
- spatial structure of the ecosystems.

3.2.1. Assessment of community structures

Ecosystems biodiversity can be observed at different levels, depending on the kind of observations. One usually distinguish:

1. *The taxonomic diversity.* Several techniques developed by molecular biologists can gather information on the genetic structure of communities:
   - *sequencing of a given gene in the community.* The RNA 16S gene is often chosen to identify bacteria or Archeae.
   - *molecular fingerprints.* Some regions in the sequence of the RNA 16S gene encode faithfully the taxa species and can be amplified by PCR techniques.
   - *the sequencing of the overall genetic material of a community* (meta-genomic)

   All these techniques bring new problems of data interpretation to estimate in a robust manner the properties of communities. The signals are combinations of contributions of abundances from each taxon. For an ecosystem with a limited diversity, composed of known species, the signal allows to determine with no ambiguity the abundances. In natural ecosystems, the signal is more complex and it is hopeless to determine uniquely the taxa distribution.
2. *The functional diversity.* It is usually observed at a larger scale, measuring the performances of the overall ecosystem to convert organic matter. The taxonomic diversity does not usually provide such information (it is possible to study *functional genes* but this is much more difficult than studying the 16S one).

A convenient way to study the functional performance of microbial community dynamics is to grow the same microbial community on different substrate compositions, and monitor its performance on these different substrates. Neutral community models [82] provide a reference for what would happen if no functional differences are present in the community. The deviation of experimental observations from neutral model predictions can be considered as a measure of functional diversity.

Understanding the links between taxonomic and functional diversity is currently a tremendous research question in biology about genotype/phenotype links, that one can also find in the specific context of microbial ecosystems.

3.2.2. Characterization of the interactions

The role of biodiversity and its preservation in ecosystems are research questions currently largely open in ecology. The nature and number of interactions between bacterial populations are poorly known, and are most probably a key to understand biodiversity. In the classical chemostat model, inter-specific interactions are rarely considered. Also in theoretical ecology, interaction information is typically encoded in an *interaction matrix*, but the coupling with common abiotic resources and the stoichiometry is hardly addressed in the models.

The information provided by confocal microscopy is also a way to estimate the distance of interactions between microorganisms and substrates. This knowledge is not often documented although it is crucial for the construction of IBM.

3.2.3. Observation of spatial structures

Schematically, one can distinguish two origins of spatialization:

1. due the physics of the environment. In bioprocesses, this happens typically for large tank size (inducing *dead zones*) or sludge accumulation making the suspension closer from a porous medium than a liquid one. Numerical experimentation can be driven, coupling a solver of the equations of the fluid mechanics with microbiology equations. Then, the spatial distribution of the biomass can be observed and used to calibrate simpler models. Typically, a dead zone is modelled as a diffusive interconnection between two perfect (abstract) tanks.

   But the biotechnology industry aims at considering more sophisticated devices than simple tanks. For instance, the fluidized bed technique consists in creating a counter-current with oxygen bubbles for preventing the biomass to leave the rector. In more complex systems, such as soil ecosystems, it is difficult to obtain faithful simulations because the spatial structure is rarely known with accuracy. Nevertheless, local observations at the level of pores can be achieved, providing information for the construction of models.
2. due to the formation of aggregates (flocks, biofilms...) or biomass wall attachment. Patterns (from ten to a hundred micro-meters) can be observed with confocal microscopy.

   Spatial distribution of bacteria, shape of patterns and composition of the aggregates help to express hypotheses on individual behaviors. But quantification and variability of images provided by confocal microscopy are difficult. An open question is to determine the relevant morphological indicators that characterize aggregation and the formation of biofilms.

3.3. Identifying, controlling and optimizing bioprocesses

The dynamics of the microbial models possess specificities that do not allow the application of the popular methods of the theory of automatic control [71], such as linear control, feedback linearization or canonical forms.

- positivity constraints. State variables, as well as control inputs, have to stay non-negative (input flow pump cannot be reversed because of contamination issues).
- non-linearity. Several models have non-controllable or non-observable linearizations when inhibition effects are present (i.e. change of monotonicity in the growth curves).
- model and measurement uncertainties. In biology, it is rarely relevant to consider model uncertainties as additive Gaussian or finite energy signals.

3.3.1. Software sensors and identification

Sensors in biology are often poor and do not provide the measurements of all the state variables of the models: substrate, strain and product concentrations. In addition, measurements are often spoilt by errors. For instance optical density measurements give an indirect measure of the biomass, influenced by abiotic factors that share the same medium.

Analytical techniques are well suited to ODE models of small dimension, such as:

- guaranteed set-membership observers, when the system is non observable or in presence of unknown inputs,
- (non-linear) changes of coordinates, when the system is observable but not in a canonical form for the construction of observers with exponential convergence.

Software sensors can be also derived with the help of simulation based approaches like particle filtering techniques [75], [74]. This method is suited to diffusion models that approximate birth and death processes. Such softwares will allow us to investigate the different sources of randomness: demography, environment, but mainly imprecision of the sensors.

Similarly, identification techniques for constant parameters are based on sensor models as well as demography and environmental randomness models. In this case, Bayesian and non-Bayesian statistical techniques can be used [73], [70].

3.3.2. Bioprocess stabilization

In bioprocesses, the most efficient bacterial species at steady state are often inhibited by too large concentrations of substrates (this corresponds to assuming that the growth function $s \mapsto \mu(s)$ in the classical chemostat model is non-monotonic). This implies that the washout equilibrium (i.e. disappearance of the biomass) can be attractive, making the bioprocess bi-stable.

A common way to globally stabilize the dynamics toward the efficient equilibrium is to manipulate the dilution rate D [71]. But a diminution of the input flow rate for the stabilization requires to have enough room for an upstream storage, which is an expensive solution especially for developing countries that need to be equipped with new installations.

Alternative ways are proposed to stabilize bioprocesses without restricting the input flow rate:

- either by *physical means*, in terms of recirculation and bypass loops, or membranes as a selective way to keep bacteria and their aggregates inside the tank and improve its efficiency.
- either by *biological means*. The *biological control* consists in adding a small quantity another species with particular growth characteristics, that will help the other species to win the competition in the end.

3.3.3. Optimal control of bioreactors

The filling stage of bioreactors, or "fed-batch", is often time consuming because the quantity of initial biomass is small and consequently the population growth is slow. The minimal time is a typical criterion for designing a filling strategy, but the optimal feedback synthesis is non trivial and may present singular arcs when the growth function is non-monotonic [86].

Recent progress have been made in the consideration of

- multi-species in sequential reactors (having more than one strain makes significantly more difficult to analyze singular arcs because of the higher dimensions of the state space, and there is little literature on the subject),
- energy consumption of flow pumps and the value of byproducts of the biological reactions such as biogas in the criterion (instead of minimal time or as penalties). Recent concerns about sustainable development encourage engineers to look for compromises between those objectives under constraints on output concentrations.

3.3.4. Plant design and optimization

We distinguish two kind of setups:

1. *The industrial setup.* A research question, largely open today, is to identify networks of interconnections of bioreactors that are the most relevant for industrial applications in terms of the following objectives:
   - reasonably simple configurations (i.e. with a limited number of tanks and connections),
   - significant improvement of the residence time at steady state over single or simpler configurations, or shapes of the reservoirs such that the total volume required for a given desired conversion factor at steady state is reduced.
2. *The bioremediation setup.* Typically, the concentration of pollutant in a natural reservoir is solution of a transport-diffusion PDE, but the optimal control of the transport term is almost not studied in the literature.

MOISE Project-Team

3. Scientific Foundations

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 Team

3. Scientific Foundations

3.1. Scientific Foundations

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 imag- ing 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, $\cdots$) and the unavoidable defects of in vivo imaging (artifacts, missing data, $\cdots$). 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, $\cdots$) 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.

NEUROMATHCOMP Project-Team

3. Scientific Foundations

3.1. Neural networks dynamics

The study of neural networks is certainly motivated by the dream 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 conductance-based Integrate and Fire models [2], [64], [65], models of epilepsy [77], effects of synaptic plasticity (see the corresponding section below).
Selected publications on this topic.

3.2. Mean-field approaches

Modelling 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 those equations. 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. Jansen-Rit, 95) 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 the DYVA team at INT, 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 integrodifferential 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 different time scales and of noise.
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 hidden probability 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) [3],[13], [63] . 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 [43], [12]. On the other hand we are proposing new algorithms for data processing [22]. We have developed a C++ software for spike train statistics based on Gibbs distributions analysis and freely available at http://enas.gforge.inria.fr/v3/. We are using this software in collaboration with several biologist groups involved in the analysis of retina spike trains (Centro de Neurociencia Valparaiso; Molecular Biology Lab, Princeton; Institut de la vision, Paris) [26], [22], [55].
Selected publications on this topic.

3.5. Synaptic Plasticity

Neural networks show amazing abilities for information storage and processing, and stimulus-dependent activity shaping, to evolve and adapt. 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 behavioural states (e.g. transition between sleep and wakeful activity). Therefore, understanding the effects of synaptic plasticity on neurons dynamics is a crucial challenge. On experimental grounds, different synaptic plasticity mechanisms have been exhibited from the Hebbian's ones (Hebb, 1949) to Long Term Potentiation (LTP) and Long Term Depression (LTD), and more recently to Spike Time Dependent Plasticity (STDP) (Markram, Lubke, Frotscher, and Sakmann, 1997; Bi and Poo, 2001). Synaptic plasticity implies that activity guides the way synapses evolve; but the resulting connectivity structure in turn can raise new dynamical regimes. This interaction becomes even more complex if the considered basic architecture is not feed-forward but includes recurrent synaptic links, like in cortical structures. Understanding this mutual coupling between dynamics and topology and its effects on the computations made by the network is a key problem in computational neuroscience.

Our group is developing mathematical and numerical methods to analyse this mutual interaction. Especially, 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 [73], [74],[4]. Also, we are currently studying the conjugated effects of synaptic and intrinsic plasticity in collaboration with H. Berry (Inria Beagle) and B. Delord, J. Naudé, ISIR team, Paris.
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 mainly focused on the study of retinal processing and motion integration at the level of V1 and MT cortical areas.

At the retina level, we are modelling its circuitry [6] and we are studying the statistics of the spike train output (see, e.g., the software ENAS http://enas.gforge.inria.fr/v3/). Real cell recordings are also analysed in collaboration with Institut de la Vision (Paris). At the level of V1-MT cortical areas, we are investigating the temporal dynamics of motion integration for a wide range of visual stimuli [38], [41], [23], [39], [42] [75], [62][5]. This work is done in collaboration with Institut de Neuroscience 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 [76], coding/decoding approaches [20], [21], [37] [69], [68] or classification [14] [66].

Selected publications on this topic.

NUMED Project-Team

3. Scientific Foundations

3.1. Multiscale modeling and computations

3.1.1. Spatial complexity: collective motion of cells

The collective motion of cells (bacteries on a gel or endothelial cells during angiogenesis) is a fascinating subject, that involves a combinaison of random walk and chemotaxis. The modeling of these problems is still active, since the pioneering works of Keller and Segel, and the mathematical study of the arising equations is a very active area of research.

Vincent Calvez focuses its effort on the following questions:

- Mathematical analysis of the Keller-Segel model

  [In collaboration with J.A. Carrillo and J. Rosado (UAB, Barcelona)]

  Following McCann 1997 and Otto 2001, we interpret the classical Keller-Segel system for chemotaxis as a gradient flow in the Wasserstein space. The free-energy functional turns out to be homogeneous. This viewpoint helps to understand better blow-up mechanisms, and to derive rates of convergence towards self-similar profiles. We investigate more precisely linear diffusion, porous medium diffusion and fast diffusion in competition with various interaction kernels.

  [In collaboration with N. Meunier (Paris 5) and R. Voituriez (Paris 6)]

  Another project consists in analyzing some variant of the Keller-Segel system when the chemoattractant is secreted at the boundary of the domain. This is motivated by modeling issues in cell polarization.

- Kinetic models for bacterial collective motion

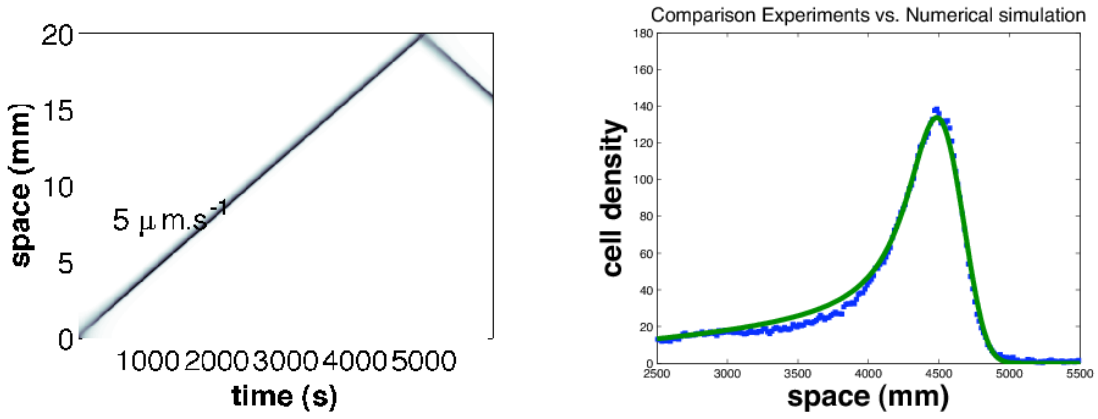


Figure 1. (left) Numerical simulation of a traveling pulse obtained with the kinetic model (right) Comparison between the bacteria density measured experimentally (blue dots) and the density computed from the kinetic model.

We have investigated kinetic models for bacterial chemotaxis following Alt and co-authors, Erban and Othmer, Dolak and Schmeiser.

We have developped a quantitative approach based on a couple of experiments performed by J. Saragosti in the team of A. Buguin and P. Silberzan (Institut Curie, Paris). These experiments describe with full statistical details solitary waves of bacteria E. coli in narrow channels. On the first set of experiments we have demonstrated that the drift-diffusion approximation of the kinetic model is valid and it fits the data very well (publication in PLoS Comput. Biol. 2010). On the second set of experiments we have simulated the kinetic model to obtain the best results as compared to the data (Fig. 1) (publication in PNAS 2011). Interestingly enough, the collaboration has led to the first experimental evidence of directional persistence of *E. coli* (the deviation angle after tumbling is smaller when the trajectory before tumbling goes in a favorable direction). We have demonstrated that this "microscopic effect" has a significant macroscopic influence on the solitary wave ($+30\%$ for the speed of the wave).

Based on these encouraging results, we have started a synthetic analysis of hyperbolic equations for chemotaxis and traveling waves.

In collaboration with Ch. Schmeiser (Univ. Vienna) we have investigated a simple (linear) kinetic equation for bacterial chemotaxis. We have obtained the existence of a stationary cluster (stable density distribution). We aim at applying the hypocoercivity results of Dolbeault-Mouhot-Schmeiser to derive a quantitative speed of relaxation towards the stable configuration. This work is under finalization.

In collaboration with N. Bournaveas (Univ. Edinburgh), C. di Russo (Univ. Lyon 1) and M. Ribot (Univ. Nice Sophia-Antipolis) we are studying hyperbolic models for cell motion. We improve the results obtained by Natalini-di Russo. These models are preliminary models which are to be complexified in order to describe growth of biofilms. This work is under progress.

In collaboration with E. Bouin and G. Nadin, we are analysing traveling waves arising in kinetic-growth equations. Namely, we study the coupling between a simple kinetic BGK operator (relaxation towards a given Maxwellian) and a logistic growth term. We have improved earlier results by Gallay-Raugel and Fedotov concerning the one-dimensional case with only two velocities. This work has been submitted. We continue the analysis with the full BGK operator. Counter-intuitive results have to be investigated further.

3.1.2. Modeling of spontaneous cell polarization

We have analysed recent models describing spontaneous polarization of cells (e.g. neuron growth cones or budding yeast). These models combine a diffusive term (in the cytoplasm) plus an advective field created at the membrane and diffusing in the cytoplasm (accounting for the actin network or the microtubules). This can be compared to the classical Keller-Segel model where diffusion competes with a non-local attractive field. Going beyond linear stability analysis we have used our know-how of the Keller-Segel system to derive global existence (no polarization) and blow-up (possibly polarization) criteria. We have also performed some numerical experiments to determine the models which exhibit spontaneous polarization. We have confirmed the prediction made by the physicists claiming that the microtubules cannot drive the cell into spontaneous polarization whereas the actin network can (Fig. 2).

Preliminaries results have been published in CRAS 2010 and SIAM J. Appl. Math (in press). We continue this project towards comparison with experimental data obtained in Matthieu Piel's lab at Institut Curie. A secondary goal consists in deriving a mechanistic model for the growth of the fission yeast Pombe. This is an ongoing work with A. Boudaoud (ENS de Lyon), N. Meunier (Univ. Paris 5), M. Piel (Institut Curie), P. Vigneaux (ENS de Lyon) and R. Voituriez (Univ. Paris 6). This is part of an ANR project JCJC, named "MODPOL" (Jan. 2012 – Dec. 2014). The project is coordinated by V. Calvez. It involves Th. Lepoutre (Inria Dracula), N. Meunier (Univ. Paris 5), M. Piel (Institut Curie), P. Vigneaux (ENS de Lyon) and R. Voituriez (Univ. Paris 6).

3.1.3. Polymerization-fragmentation processes

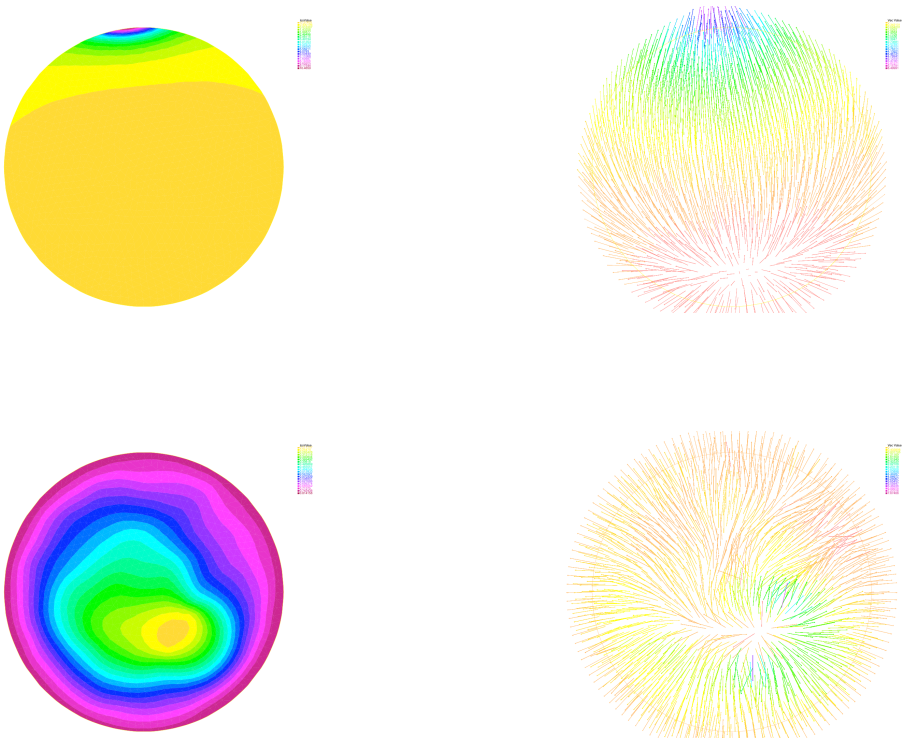


Figure 2. 2D numerical simulations of cell polarization on a round shaped cell. (Top) The actin network carries the attractive field: polarization occurs. (Bottom) The microtubules carry the attractive field: we observe no polarization. (Work in progress; simulations are done with FreeFEM++)

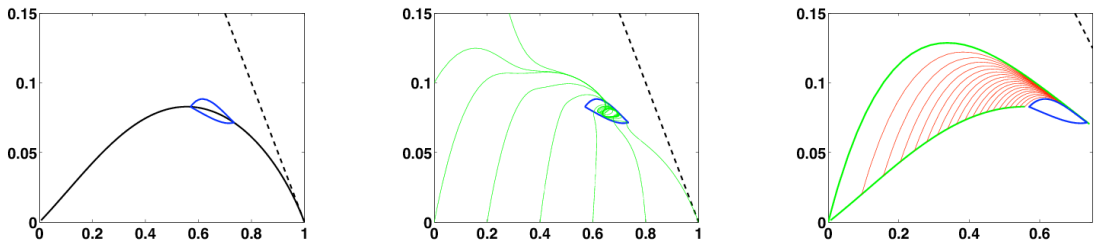


Figure 3. Dynamics of trajectories of the control system projected on the simplex. (left) Remarkable sets in the simplex: the line of eigenvectors parametrized by the control parameter, and the small "ergodic" set. (middle) All trajectories eventually enter the ergodic set. (right) We prove a tunelling effect: all trajectories are confined in a neighbourhood of the ergodic set, and moves towards it.

In collaboration with M. Doumic (Inria Bang) and P. Gabriel (Inria Beagle) we have studied the behaviour of the eigenvalue problem for genuine growth-fragmentation equations. We have focused on the dependence of the couple eigenvalue-eigenvector with respect to the growth and fragmentation coefficients. We have mainly used blowing-techniques and asymptotic estimates. We have shown counter-intuitive (non-monotonic) dependence. We have also discussed the possible consequences on applications.

Together with P. Gabriel (Inria Beagle) we are investigating the optimal control problem for a baby polymerization-fragmentation process mimicking the controlled growth of PrPres (prion) polymers. It consists in a three compartments system (small, intermediate and large polymers) with linear transitions between the compartments. We have a single control parameter acting on the fragmentation process.

We first assume that the control parameter has to be chosen constant. Under certain conditions, there is a best possible choice with infinite-time horizon. It maximizes the exponential growth by optimizing the eigenvalue of the polymerization-fragmentation matrix.

When we relax the condition of constant control, we have to deal with an optimal control problem. It can be translated into a Hamilton-Jacobi-Bellman equation. Although it is a very degenerated case, we can prove existence and uniqueness of an infinite-horizon eigenvalue, as in the constant case. We use the notion of ergodic set introduced by Arisawa-Lions (1998). The success of the proof relies on refined analysis of the dynamics of close-to-optimal trajectories projected on the simplex (Fig. 3). This work is under finalization.

3.1.4. Complex rheology

To investigate the growth of a tumor it is crucial to have a correct description of its mechanical aspects. Tumoral and normal cells may be seen as a complex fluid, with complex rheology.

Numerical investigations of complex flows is studied by P. Vigneaux who develops new numerical schemes fo Bingham type flows.

3.2. Parametrization of complex systems

The parametrization of complex systems in order to fit experimental results or to have a good qualitative behavior is a delicate issue since its require to simulate the complex systems for a large number of sets of parameters, which is very expensive.

In many medical contexts, the available data for one particular patient are rather poor (a few MRI for instance). However many patients are studied (20 to 100 or even more in frequent pathologies). Therefore it is difficult or even impossible to parametrize a model for a given patient (too many parameters with respect to the number of available clinical data). However, it is possible to infer the distribution of the parameters in the global population by using all the data of all the patients at the same time. This is the principle of populational parametrization: to look for the distribution of the parameters (Gaussian or log Gaussian) and not to try to study each patient individually.

Many algorithms have been developed for populational parametrization, in particular so called SAEM (Stochastic Approximation Expectation Maximization) algorithms, based on MCMC (Monte Carlo Markov Chain) algorithms. These algorithms are very expensive, and require hundreds of thousands of evaluations of the model. For ordinary differential equation based models, SAEM converge quickly (it takes ten to twenty minutes on a laptop for the Monolix implementation of SAEM. Monolix is developped by M. Lavielle at Inria).

However for PDE based models, the evaluation of one single model may be long (a few minutes, up to ten minutes), hence the evaluation of hundreds of thousands models is completely out of range. Moreover, SAEM can not be parallelized in an efficient way.

Numed has set a general strategy to allow populational approaches on complex systems or on PDE based models. It relies on a precomputation strategy, combined iteratively with SAEM algorithms.

With such a strategy, populational parametrization of a PDE like reaction diffusion equation (KPP) may be done on a few hours on a small cluster of cores (32 cores).

PARIETAL Project-Team

3. Scientific Foundations

3.1. Human neuroimaging data and its use

Human neuroimaging consists in acquiring non-invasively image data from normal and diseased human populations. Magnetic Resonance Imaging (MRI) can be used to acquire information on brain structure and function at high spatial resolution.

- T1-weighted MRI is used to obtain a segmentation of the brain into different different tissues, such as gray matter, white matter, deep nuclei, cerebro-spinal fluid, at the millimeter or sub-millimeter resolution. This can then be used to derive geometric and anatomical information on the brain, e.g. cortical thickness.
- Diffusion-weighted MRI measures the local diffusion of water molecules in the brain at the resolution of 2mm, in a set of directions (30 to 60 typically). Local anisotropy, observed in white matter, yields a geometric model of fiber tracts along which water diffusion occurs, and thus provides essential information of the connectivity structure of the brain.
- Functional MRI measures the blood-oxygen-level-dependent (BOLD) contrast that reflects neural activity in the brain, at a spatial resolution of 2 to 3mm, and a temporal resolution of 2-3s. This yields a spatially resolved image of brain functional networks that can be modulated either by specific cognitive tasks or appear as networks of correlated activity.
- Electro- and Magneto-encephalography (MEEG) are two additional modalities that complement functional MRI, as they directly measure the electric and magnetic signals elicited by neural activity, at the millisecond scale. These modalities rely on surface measurements and do not localize brain activity very accurately in the spatial domain.

3.2. High-field MRI

High field MRI as performed at Neurospin (7T on humans, 11.7T in 2013, 17.6T on rats) brings an improvement over traditional MRI acquisitions at 1.5T or 3T, related to to a higher signal-to-noise ratio in the data. Depending on the data and applicative context, this gain in SNR can be traded against spatial resolution improvements, thus helping in getting more detailed views of brain structure and function. This comes at the risk of higher susceptibility distortions of the MRI scans and signal inhomogeneities, that need to be corrected for. Improvements at the acquisition level may come from the use of new coils (such as the new 32 channels coil on the 7T at Neurospin).

3.3. Technical challenges for the analysis of neuroimaging data

The first limitation of Neuroimaging-based brain analysis is the limited Signal-to-Noise Ratio of the data. A particularly striking case if functional MRI, where only a fraction of the data is actually understood, and from which it is impossible to observe by eye the effect of neural activation on the raw data. Moreover, far from traditional i.i.d. Gaussian models, the noise in MRI typically exhibits correlations and long-distance correlation properties (e.g. motion-related signal) and has potentially large amplitude, which can make it hard to distinguish from true signal on a purely statistical basis. A related difficulty is the *lack of salient structure* in the data: it is hard to infer meaningful patterns (either through segmentation or factorization procedures) based on the data only. A typical case is the inference of brain networks from resting-state functional connectivity data.

Regarding statistical methodology, neuroimaging problems also suffer from the relative paucity of the data, i.e. the relatively small number of images available to learn brain features or models, e.g. with respect to the size of the images or the number of potential structures of interest. This leads to several kinds of difficulties, known either as multiple comparison problems or curse of dimensionality. One possibility to overcome this challenge is to increase the amount of data by using images from multiple acquisition centers, at the risk of introducing scanner-related variability, thus challenging the homogeneity of the data. This becomes an important concern with the advent of cross-modal neuroimaging-genetics studies.

POMDAPI Project-Team (section vide)

REO Project-Team

3. Scientific Foundations

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.

Because of the above mentioned difficulties, the interaction between the blood flow and the artery wall has often been neglected in most of the classical studies. The numerical properties of the fluid-structure coupling in blood flows are rather different from other classical fluid-structure problems. In particular, due to stability reasons it seems impossible to successfully apply the explicit coupling schemes used in 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 over the last few years several efficient fluid-structure interaction algorithms. We are now using these algorithms to address inverse problems in blood flows (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 [77] in the frame of combustion. It was later used to develop the Kiva code [67] at the Los Alamos National Laboratory, or the Hesione code [72], for example. It has a wide range of applications, besides the nuclear setting: diesel engines, rocket engines [70], 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 [73]. 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 [1]. 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 plan to mix arbitrary Lagrangian Eulerian and fictitious domain approaches, or simplified valve models based on an immersed surface strategy.

[1] 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 [76], and Biot [68] 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 [69]), 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 to develop a multiscale modeling 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. Scientific Foundations

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 Team

3. Scientific Foundations

3.1. Glossary

WF Optical Wide-Field microscopy.

SDC (Spinning-Disk Confocal microscopy): illumination of the sample with a rotating pattern of several hundred of pinholes for complete simultaneous confocal illumination.

FLIM (Fluorescence Lifetime Microscopy Imaging): imaging of fluorescent molecule lifetimes.

PALM (Photo-Activated Localization Microscopy): high-resolution microscopy using stochastic photo-activation of fluorophores and adjustment of point spread functions [20].

SIM (Structured Illumination Microscopy): high-resolution light microscopy using structured patterns and interference analysis [30].

TIRF (Total Internal Reflectance): 2D optical microscopy using evanescent waves and total reflectance [19].

Cryo-EM (Cryo-Electron Tomography): 3D representation of sub-cellular and molecular objects of 5-20 nanometres, frozen at very low temperatures, from 2D projections using a transmission electron microscope.

3.2. Image restoration for high-resolution microscopy

In order to produce images compatible with the dynamic processes in living cells as seen in video-microscopy, we study the potential of non-local neighborhood filters and image denoising algorithms (e.g. ND-SAFIR software) [6], [2], [7], [4]. The major advantage of these approaches is to acquire images at very low SNR while recovering denoised 2D+T(ime) and 3D+T(ime) images [1]. Such post-acquisition processing can improve the rate of image acquisition by a factor of 100 to 1000 times [5], reducing the sensitivity threshold and allowing imaging for long time regime without cytotoxic effect and photodamages. This approach has been successfully applied to WF, SDC [1], TIRF [19], fast live imaging and 3D-PALM using the OMX system in collaboration with J. Sedat and M. Gustafsson at UCSF [5]. The ND-SAFIR software (see Section 5.1) has been licensed to a large set of laboratories over the world (see Figure 2). 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 [32], PALM [20]) SIM has the strong advantage of versatility when considering the photo-physical properties of the fluorescent probes [30]. 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.

3.3. Dynamic analysis and trajectory computation

3.3.1. Motion analysis and tracking

In time-lapse microscopy, the challenge is to detect and track moving objects. 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 [37]. Data association even combined with sophisticated particle filtering techniques [40] or matching techniques [38] is problematic when tracking several hundreds of similar objects with variable velocities. Developing new optical flow and 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. The goal is to formulate the problem of optical flow estimations in ways that take physical causes of brightness violations into account [26], [31]. In addition, the interpretation of computed flow fields enables to provide spatio-temporal signatures of particular dynamic processes and could help to complete the traffic modelling.

3.3.2. Event detection

Several approaches can be considered for the automatic detection of appearing and vanishing particles (or spots) in WF and TIRF microscopy images. The difficulty is to distinguish motions due to trafficking from the appearing and vanishing spots. Ideally this could be performed by tracking all the vesicles contained in the cell [40], [29]. Among the methods proposed to detect particles in microscopy images [43], [39], 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 [24]).

3.4. Computational 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 [33], [45] 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 [35] 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) [44] 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 will be to correlate stochastic and dynamical 1D and in silico models studied at the nano-scale in biophysics, to 3D images acquired in vivo at the scale of few hundred nanometres. A difficulty is related to the scale change and statistical aggregation problems (in time and space).

SHACRA Project-Team

3. Scientific Foundations

3.1. Biomechanical Modeling

3.1.1. 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 such an approach [24], [27] 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 rely on linearized constitutive laws.

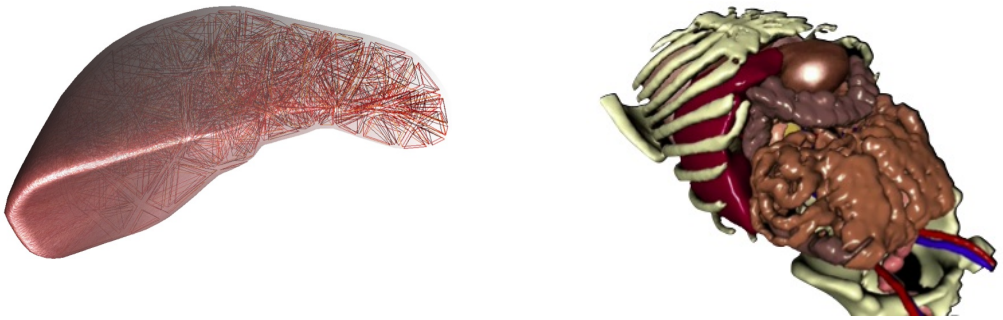


Figure 1. Biomechanical models of organs, based on the Finite Element Method and elasticity theory. Left: a model of the liver based on tetrahedral elements and small strain elasticity. Right: several organ models from a patient dataset combined to create a realistic abdominal anatomy.

An important part of our research is dedicated to the development of new, more accurate models that remain 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. Yet, patient-specific planning or per-operative guidance also requires the models to be parametrized with patient-specific biomechanical data. Very little work has been done in this area, in particular when tissue properties need to be measured in vivo non-invasively. New imaging techniques, such as Ultrasound Elastography or Magnetic Resonance Elastography, could be used to this end [23]. We are currently studying the impact of parametrized patient-specific models of the liver in the context of the PASSPORT european project. This will be used to provide information about the deformation, tissue stiffness and tumor location, for various liver pathologies.

3.1.2. Biomechanical modeling of hollow structures

A large number of anatomical structures in the human body are vascularized (brain, liver, heart, kidneys, ...) and recent interventions (such as interventional radiology) rely on the vascular network as a therapeutical pathway. It is therefore essential to model the shape and deformable behavior of blood vessels. This will be

done at two levels. Global deformation of a vascular network: we have demonstrated previously [9] that we could recover the shape of thousands of vessels from medical images by extracting the centerline of each vessel (see Figure 2). The resulting vascular skeleton can be modeled as a deformable (tree) structure which can capture the global aspects of the deformation. More local deformations can then be described by considering now the actual local shape of the vessel. Other structures such as aneurysms, the colon or stomach can also benefit from being modeled as deformable structures. For this we will rely on shell or thin plate theory. We have recently obtained very encouraging results in the context of the Ph.D. thesis of Olivier Comas [26]. Such local and global models of hollow structures will be particularly relevant for planning coil deployment or stent placement, but also in the context of a new laparoscopic technique called NOTES which uses a combination of a flexible endoscope and flexible instruments. Obtaining patient-specific models of vascular structures and associated pathologies remains a challenge from an image processing stand point, and this challenge is even greater once we require these models to be adapted to complex computational strategies. To this extend we will pursue our collaboration with the MAGRIT team at Inria (through a PhD thesis starting in January 2010) and the Massachusetts General Hospital in Boston.

3.1.3. Blood Flow Simulation

Beyond biomechanical modeling of soft tissues, an essential component of a simulation is the modeling of the functional interactions occurring between the different elements of the anatomy. This involves for instance modeling physiological flows (blood flow, air flow within the lungs...). We particularly plan to study the problem of fluid flow in the context of vascular interventions, such as the simulation of three-dimensional turbulent flow around aneurysms to better model coil embolization procedures. Blood flow dynamics is starting to play an increasingly important role in the assessment of vascular pathologies, as well as in the evaluation of pre- and post-operative status. While angiography has been an integral part of interventional radiology procedures for years, it is only recently that detailed analysis of blood flow patterns has been studied as a mean to assess complex procedures, such as coil deployment. A few studies have focused on aneurysm-related hemodynamics before and after endovascular coil embolization. Groden et al. [31] constructed a simple geometrical model to approximate an actual aneurysm, and evaluated the impact of different levels of coil packing on the flow and wall pressure by solving Navier-Stokes equations, while Kakalis et al. [33] relied on patient-specific data to get more realistic flow patterns, and modeled the coiled aneurysm as a porous medium. As these studies aimed at accurate Computational Fluid Dynamics simulation, they rely on commercial software, and the computation times (dozens of hours in general) are incompatible with interactive simulation or even clinical practice. Generally speaking, accuracy and efficiency are two significant pursuits in numerical calculation, but unfortunately very often contradictory.

With the Ph.D. thesis of Yiyi Wei, we have recently started the development of a new technique for accurately computing, in near real-time, the flow of blood within an aneurysm, as well as the interaction between blood and coils. In this approach we rely on the Discrete Exterior Calculus method to obtain an ideal trade-off between accuracy and computational efficiency. Although still at an early stage, these results show that our approach can accurately capture the main characteristics of the complex blood flow patterns in and around an aneurism. The model also takes into account the influence of the coil on the blood flow within the aneurysm. The main difference between our approach and many other work done by internationally renowned teams (such as REO team at Inria or the Computer Vision Laboratory at ETH) comes from the importance we place in the computational efficiency of the method. To some extent our approach is similar to what has been done to obtain real-time finite element methods. We are essentially trying to capture the key characteristics of the behavior for a particular application. This is well illustrated by the work we started on flow modeling, which received an award in September 2009 at the selective conference on Medical Image Computing and Computer Assisted Interventions [10]. We will pursue this direction to accurately model the local flow in a closed domain (blood vessel, aneurysm ventricle, ...) and combine it with some of our previous work describing laminar flow across a large number of vessels [38] in order to define boundary conditions for the three-dimensional model.

3.2. Biomechanical Systems

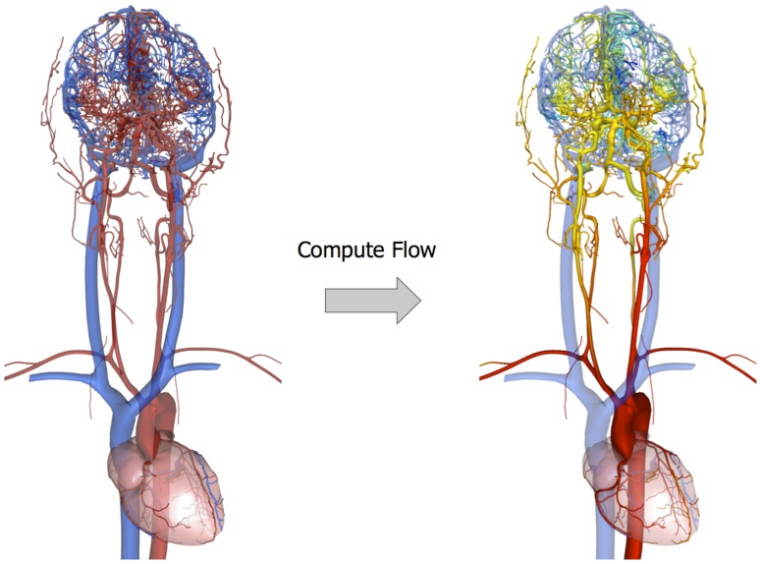


Figure 2. Blood flow and pressure distribution in the cerebrovascular system. The arterial vascular network is composed of more than 3,000 vessels, yet the computation is performed in real-time.

3.2.1. Constraint models and boundary conditions

To accurately model soft tissue deformations, the approach must account for the intrinsic behavior of the target organ, but also for its biomechanical interactions with surrounding tissues or with medical devices. While the biomechanical behavior of important organs (such as the brain or liver) has been well studied, few work exists regarding the mechanical interactions between the anatomical structures. For tissue-tool interactions, most approaches rely on a simple contact models, and rarely account for friction. While this simplification can produce plausible results in the case of an interaction between the end effector of a laparoscopic instrument and the surface of an organ, it is generally an incorrect approximation. As we move towards simulations for planning or rehearsal, accurately modeling contacts will take an increasingly important place. We have recently shown in [28] and [29] that we could compute, in real-time, complex interactions between a coil and an aneurysm, or between a flexible needle and soft-tissues. In laparoscopic surgery, the main challenge lies in the modeling of interactions between anatomical structures rather than between the instruments and the surface of an organ. During the different steps of a procedure organs slides against each other, while respiratory, cardiac and patient motion also generate contacts. Modeling these multiple interactions becomes even more complex when different biomechanical models are used to characterize the various soft tissues of the anatomy. Consequently, our objective is to accurately model resting contacts with friction, in a heterogeneous environment (spring-mass models, finite element models, particle systems, rigid objects, etc.). When different time integration strategies are used, a challenge lies in the computation of contact forces in a way that integrity and stability of the overall simulation are maintained. Our objective is to work on the definition of these various boundary conditions and on new resolution methods for such heterogeneous simulations. In particular we will investigate a simulation process in which each model continues to benefit from its own optimizations while taking into account the mechanical couplings due to interactions between objects.

3.2.2. Vascularized anatomy

From a clinical standpoint, several procedures involve vascularized anatomical structures such as the liver, the kidneys, or the brain. When a therapy needs to be applied on such structures, it is currently possible to perform

a procedure surgically or to use an endovascular approach. This requires to characterize and model the behavior of vessels (arteries and veins) as well as the behavior of soft tissue (in particular the parenchyma). Another challenge of this research will be to model the interactions between the vascular network and the parenchyma where it is embedded. These interactions are key for both laparoscopic surgery and interventional radiology as they allow to describe the motion of the vessels in a vascularized organ during the procedure. This motion is either induced by the surgical manipulation of the parenchymal tissue during surgery or by respiratory, cardiac or patient motion during interventional radiology procedures. From a biomechanical standpoint, capillaries are responsible for the viscoelastic behavior of the vascularized structures, while larger vessels have a direct impact on the overall behavior of the anatomy. In the liver for instance, the apparent stiffness of the organ changes depending on the presence or absence of large vessels. Also, the relatively isotropic nature of the parenchyma is modified around blood vessels. We propose to model the coupling that exists between these two different anatomical structures to account for their respective influence. For this we will initially rely on the work done during the Ph.D. thesis of Christophe Guebert (see ([32] for instance) and we will also investigate coupling strategies based on degrees of freedom reduction to reduce the complexity of the problem (and therefore also computation times). Part of this work is already underway in the context of the PASSPORT european project with IRCAD and soft tissue measurements will be performed in collaboration with the biomechanics laboratory at Strasbourg University.

3.2.3. Parallel Computation

Although the past decade has seen a significant increase in complexity and performance of the algorithms used in medical simulation, major improvements are still required to enable patient-specific simulation and planning. Using parallel architectures to push the complexity of simulated environments further is clearly an approach to consider. However, interactive simulations introduce new constraints and evaluation criteria, such as latencies, multiple update frequencies and dynamic adaptation of precision levels, which require further investigation. New parallel architectures, such as multi-cores CPUs, are now ubiquitous as the performances achieved by sequential units (single core CPUs) stopped to regularly improve. At the same time, graphical processors (GPU) offer a massive computing power that is now accessible to non-graphical tasks thanks to new general-purposes API such as CUDA and OpenCL. GPUs are internally parallel processors, exploiting hundreds of computing units. These architectures can be exploited for more ambitious simulations, as we already have demonstrated in a first step by adding support for CUDA within the SOFA framework. Several preliminary results of GPU-based simulations have been obtained, permitting to reach speedup factors (compared to a single core GPU) ranging from 16x to 55x. Such improvements permit to consider simulations with finer details, or new algorithms modeling biomechanical behaviors more precisely. However, while the fast evolution of parallel architectures is useful to increase the realism of simulations, their varieties (multi-core CPUs, GPUs, clusters, grids) make the design of parallel algorithm challenging. An important effort needs to be made is to minimize the dependency between simulation algorithms and hardware architectures, allowing the reuse of parallelization efforts on all architecture, as well as simultaneously exploiting all available computing resources present in current and future computers. The largest gains could be achieved by combining parallelism and adaptive algorithms. The design and implementation of such a system is a challenging problem, as it is no longer possible to rely on pre-computed repartition of datas and computations. Thus, further research is required in highly adaptive parallel scheduling algorithms, and highly efficient implementation able to handle both large changes in computational loads due to user interactions and multi-level algorithms, and new massively parallel architectures such as GPUs. A direction that we are also investigating is to combine multi-level representations and locally adaptive meshes. Multi-level algorithms are useful not only to speedup computations, but also to describe different characteristics of the deformation at each level. Combined with local change of details of the mesh (possibly using hierarchical structures), the simulation can reach a high level of scalability.

SISYPHE Project-Team

3. Scientific Foundations

3.1. System theory for systems modeled by ordinary differential equations

3.1.1. Identification, observation, control and diagnosis of linear and nonlinear systems

Characterizing and inferring properties and behaviors of objects or phenomena from observations using models is common to many research fields. For dynamical systems encountered in the domains of engineering and physiology, this is of practical importance for monitoring, prediction, and control. For such purposes, we consider most frequently, the following model of dynamical systems:

$$\begin{array}{rcl} \frac{dx(t)}{dt} & = & f(x(t),u(t),\theta,w(t)) \\ y(t) & = & g(x(t),u(t),\theta,v(t)) \end{array} \tag{101}$$

where $x(t)$, $u(t)$ and $y(t)$ represent respectively the state, input and output of the system, f and g characterize the state and output equations, parameterized by θ and subject to modeling and measurement uncertainties $w(t)$ and $v(t)$. Modeling is usually based on physical knowledge or on empirical experiences, strongly depending on the nature of the system. Typically only the input $u(t)$ and output $y(t)$ are directly observed by sensors. Inferring the parameters θ from available observations is known as system identification and may be useful for system monitoring [102], whereas algorithms for tracking the state trajectory $x(t)$ are called observers. The members of SISYPHE have gained important experiences in the modeling of some engineering systems and biomedical systems. The identification and observation of such systems often remain challenging because of strong nonlinearities [21]. Concerning control, robustness is an important issue, in particular to ensure various properties to all dynamical systems in some sets defined by uncertainties [83], [84]. The particularities of ensembles of connected dynamical systems raise new challenging problems.

Examples of reduced order models:
- Reduced order modeling of the cardiovascular system for signal & image processing or control applications. See section 3.3.1 .
- Excitable neuronal networks & control of the reproductive axis by the GnRH. See section 3.3.2 .
- Modeling, Control, Monitoring and Diagnosis of Depollution Systems. See section 6.1 .

3.2. System theory for quantum and quantum-like systems

3.2.1. Quantization of waves propagation in transmission-line networks & Inverse scattering

Linear stationary waves. Our main example of classical system that is interesting to see as a quantum-like system is the Telegrapher Equation, a model of transmission lines, possibly connected into a network. This is the standard model for electrical networks, where V and I are the voltage and intensity functions of z and k, the position and frequency and $R(z)$, $L(z)$, $C(z)$, $G(z)$ are the characteristics of the line:

$$\frac{\partial V(z,k)}{\partial z} = -\left(R(z)+jkL(z)\right)I(z,k), \quad \frac{\partial I(z,k)}{\partial z} = -(G(z)+jkC(z))V(z,k) \tag{102}$$

Since the work of Noordergraaf [101], this model is also used for hemodynamic networks with V and I respectively the blood pressure and flow in vessels considered as 1D media, and with $R=\frac{8\pi\eta}{S^2}$, $L=\frac{\rho}{S}$, $C=\frac{3S(r+h)}{E(2r+h)}$ where ρ and η are the density and viscosity of the blood ; r, h and E are the inner radius, thickness and Young modulus of the vessel. $S=\pi r^2$. The conductivity G is a small constant for blood flow.

Monitoring such networks is leading us to consider the following inverse problem: *get information on the functions R, L, C, G from the reflection coefficient* $\mathcal{R}(k)$ (ratio of reflected over direct waves) measured in some location by Time or Frequency Domain Reflectometry.

To study this problem it is convenient to use a Liouville transform, setting $x(z) = \int_0^z \sqrt{L(z')C(z')}dz'$, to introduce auxiliary functions $q^{\pm}(x) = \frac{1}{4}\frac{d}{x}\left(\ln\frac{L(x)}{C(x)}\right) \pm \frac{1}{2}\left(\frac{R(x)}{L(x)} - \frac{G(x)}{C(x)}\right)$ and $q_p(x) = \frac{1}{2}\left(\frac{R(x)}{L(x)} + \frac{G(x)}{C(x)}\right)$, so that (2) becomes a Zakharov-Shabat system [89] that reduces to a Schrödinger equation in the lossless case ($R = G = 0$):

$$\frac{\partial v_1}{\partial x} = (q_p - jk)\, v_1 + q^+ v_2, \quad \frac{\partial v_2}{\partial x} = -(q_p - jk)\, v_2 + q^- v_1 \tag{103}$$

and $I(x,k) = \frac{1}{\sqrt{2}}\left[\frac{C(x)}{L(x)}\right]^{\frac{1}{4}}(v_1(x,k) + v_2(x,k))$, $V(x,k) = -\frac{1}{\sqrt{2}}\left[\frac{L(x)}{C(x)}\right]^{\frac{1}{4}}(v_1(x,k) - v_2(x,k))$.
Our inverse problem becomes now an inverse scattering problem for a Zakharov-Shabat (or Schrödinger) equation: *find the potentials* $q^{\pm}$ *and* q_p *corresponding to* $\mathcal{R}$. This classical problem of mathematical physics can be solved using e.g. the Gelfand-Levitan-Marchenko method.

Nonlinear traveling waves. In some recent publications [92], [91], we use scattering theory to analyze a measured Arterial Blood Pressure (ABP) signal. Following a suggestion made in [103], a Korteweg-de Vries equation (KdV) is used as a physical model of the arterial flow during the pulse transit time. The signal analysis is based on the use of the Lax formalism: the iso-spectral property of the KdV flow allows to associate a constant spectrum to the non stationary signal. Let the non-dimensionalized KdV equation be

$$\frac{\partial y}{\partial t} - 6y\frac{\partial y}{\partial x} + \frac{\partial^3 y}{\partial x^3} = 0 \tag{104}$$

In the Lax formalism, y is associated to a Lax pair: a Schrödinger operator, $L(y) = -\frac{\partial^2}{\partial x^2} + y$ and an anti-Hermitian operator $M(y) = -4\frac{\partial^3}{\partial x^3} + 3y\frac{\partial}{\partial x} + 3\frac{\partial}{\partial x}y$. The signal y is playing here the role of the potential of $L(y)$ and is given by an operator equation equivalent to (4):

$$\frac{\partial L(y)}{\partial t} = [M(y), L(y)] \tag{105}$$

Scattering and inverse scattering transforms can be used to analyze y in term of the spectrum of $L(y)$ and conversely. The "bound states" of $L(y)$ are of particular interest: if $L(y)$ is solution of (5) and $L(y(t))$ has only bound states (no continuous spectrum), then this property is true at each time and y is a soliton of KdV. For example the arterial pulse pressure is close to a soliton [86], [14].

Inverse scattering as a generalized Fourier transform. For "pulse-shaped" signals y, meaning that $y \in L^1(\mathbb{R}; (1+|x|^2)dx)$, the squared eigenfunctions of $L(y)$ and their space derivatives are a basis in $L^1(\mathbb{R}; dx)$ (see e.g. [99]) and we use this property to analyze signals. Remark that the Fourier transform corresponds to using the basis associated with $L(0)$. The expression of a signal y in its associated basis is of particular interest. For a positive signal (as e.g. the arterial pressure), it is convenient to use $L(-y)$ as $-y$ is like a multi-well potential, and the Inverse scattering transform formula becomes:

$$y(x) = 4\sum_{n=1}^{n=N}\kappa_n\psi_n^2(x) - \frac{2i}{\pi}\int_{-\infty}^{-\infty} k\mathcal{R}(k)f^2(k,x)dk \tag{106}$$

where ψ_n and $f(k,.)$ are solutions of $L(-y)f = k^2 f$ with $k = i\kappa_n$, $\kappa_n > 0$, for ψ_n (bound states) and $k > 0$ for $f(k,.)$ (Jost solutions). The discrete part of this expression is easy to compute and provides useful informations on y in applications. The case $\mathcal{R} = 0$ ($-y$ is a reflectionless potential) is then of particular interest as $2N$ parameters are sufficient to represent the signal. We investigate in particular approximation of pulse-shaped signals by such potentials corresponding to N-solitons.

3.2.2. Identification & control of quantum systems

Interesting applications for quantum control have motivated seminal studies in such wide-ranging fields as chemistry, metrology, optical networking and computer science. In chemistry, the ability of coherent light to manipulate molecular systems at the quantum scale has been demonstrated both theoretically and experimentally [98]. In computer science, first generations of quantum logical gates (restrictive in fidelity) has been constructed using trapped ions controlled by laser fields (see e.g. the "Quantum Optics and Spectroscopy Group, Univ. Innsbruck"). All these advances and demands for more faithful algorithms for manipulating the quantum particles are driving the theoretical and experimental research towards the development of new control techniques adapted to these particular systems. A very restrictive property, particular to the quantum systems, is due to the destructive behavior of the measurement concept. One can not measure a quantum system without interfering and perturbing the system in a non-negligible manner.

Quantum decoherence (environmentally induced dissipations) is the main obstacle for improving the existing algorithms [88]. Two approaches can be considered for this aim: first, to consider more resistant systems with respect to this quantum decoherence and developing faithful methods to manipulate the system in the time constants where the decoherence can not show up (in particular one can not consider the back-action of the measurement tool on the system); second, to consider dissipative models where the decoherence is also included and to develop control designs that best confronts the dissipative effects.

In the first direction, we consider the Schrödinger equation where $\Psi(t,x)$, $-\frac{1}{2}\Delta$, V, μ and $u(t)$ respectively represent the wavefunction, the kinetic energy operator, the internal potential, the dipole moment and the laser amplitude (control field):

$$i\frac{d}{dt}\Psi(t,x) = (H_0 + u(t)H_1)\Psi(t,x) = (-\frac{1}{2}\Delta + V(x) + u(t)\mu(x))\Psi(t,x), \quad \Psi|_{t=0} = \Psi_0, \tag{107}$$

While the finite dimensional approximations ($\Psi(t) \in \mathbb{C}^N$) have been very well studied (see e.g. the works by H. Rabitz, G. Turinici, ...), the infinite dimensional case ($\Psi(t,.) \in L^2(\mathbb{R}^N;\mathbb{C})$) remains fairly open. Some partial results on the controllability and the control strategies for such kind of systems in particular test cases have already been provided [79], [80], [94]. As a first direction, in collaboration with K. Beauchard (CNRS, ENS Cachan) et J-M Coron (Paris-sud), we aim to extend the existing ideas to more general and interesting cases. We will consider in particular, the extension of the Lyapunov-based techniques developed in [95], [81], [94]. Some technical problems, like the pre-compactness of the trajectories in relevant functional spaces, seem to be the main obstacles in this direction.

In the second direction, one needs to consider dissipative models taking the decoherence phenomena into account. Such models can be presented in the density operator language. In fact, to the Schrödinger equation (7), one can associate an equation in the density operator language where $\rho = \Psi\Psi^*$ represents the projection operator on the wavefunction Ψ ($[A,B] = AB - BA$ is the commutator of the operators A and B):

$$\frac{d}{dt}\rho = -i[H_0 + u(t)H_1, \rho], \tag{108}$$

Whenever, we consider a quantum system in its environment with the quantum jumps induced by the vacuum fluctuations, we need to add the dissipative effect due to these stochastic jumps. Note that at this level, one also can consider a measurement tool as a part of the environment. The outputs being partial and not giving complete information about the state of the system (Heisenberg uncertainty principle), we consider a so-called quantum filtering equation in order to model the conditional evolution of the system. Whenever the measurement tool composes the only (or the only non-negligible) source of decoherence, this filter equation admits the following form:

$$\begin{aligned} d\rho_t = -i[H_0 + u(t)H_1, \rho_t]dt + (L\rho_t L^* - \frac{1}{2}L^*L\rho_t - \frac{1}{2}\rho_t L^* L)dt \\ +\sqrt{\eta}(L\rho_t + \rho_t L^* - \mathrm{Tr}\left[(L+L^*)\rho_t\right]\rho_t)dW_t, \end{aligned} \tag{109}$$

where L is the so-called Lindblad operator associated to the measurement, $0 < \eta \leq 1$ is the detector's efficiency and where the Wiener process W_t corresponds to the system output Y_t via the relation $dW_t = dY_t - \mathrm{Tr}\left[(L+L^*)\rho_t\right]dt$. This filter equation, initially introduced by Belavkin [82], is the quantum analogous of a Kushner-Stratonovic equation. In collaboration with H. Mabuchi and his co-workers (Physics department, Caltech), we would like to investigate the derivation and the stochastic control of such filtering equations for different settings coming from different experiments [96].

Finally, as a dual to the control problem, physicists and chemists are also interested in the parameter identification for these quantum systems. Observing different physical observables for different choices of the input $u(t)$, they hope to derive more precise information about the unknown parameters of the system being parts of the internal Hamiltonian or the dipole moment. In collaboration with C. Le Bris (Ecole des ponts and Inria), G. Turinici (Paris Dauphine and Inria), P. Rouchon (Ecole des Mines) and H. Rabitz (Chemistry department, Princeton), we would like to propose new methods coming from the systems theory and well-adapted to this particular context. A first theoretical identifiability result has been proposed [93]. Moreover, a first observer-based identification algorithm is under study.

3.3. Physiological & Clinical research topics

3.3.1. The cardiovascular system: a multiscale controlled system

Understanding the complex mechanisms involved in the cardiac pathological processes requires fundamental researches in molecular and cell biology, together with rigorous clinical evaluation protocols on the whole organ or system scales. Our objective is to contribute to to these researches by developing low-order models of the cardiac mechano-energetics and control mechanisms, for applications in model-based cardiovascular signal or image processing.

We consider intrinsic heart control mechanisms, ranging from the Starling and Treppe effects on the cell scale to the excitability of the cardiac tissue and to the control by the autonomous nervous system. They all contribute to the function of the heart in a coordinated manner that we want to analyze and assess. For this purpose, we study reduced-order models of the electro-mechanical activity of cardiac cells designed to be coupled with measures available on the organ scale (e.g. ECG and pressure signals). We study also the possibility to gain insight on the cell scale by using model-based multiscale signal processing techniques of long records of cardiovascular signals.

Here are some questions of this kind, we are considering:
- Modeling the controlled contraction/relaxation from molecular to tissue and organ scales.
- Direct and inverse modeling the electro-mechanical activity of the heart on the cell scale.
- Nonlinear spectral analysis of arterial blood pressure waveforms and application to clinical indexes.
- Modeling short-term and long-term control dynamics on the cardiovascular-system scale. Application to a Total Artificial Heart.

3.3.2. Reproductive system: follicular development & ovulation control

The ovulatory success is the main limiting factor of the whole reproductive process, so that a better understanding of ovulation control is needed both for clinical and zootechnical applications. It is necessary to improve the treatment of anovulatory infertility in women, as it can be by instance encountered in the PolyCystic Ovarian Syndrome (PCOS), whose prevalence among reproductive-age women has been estimated at up to 10%. In farm domestic species, embryo production following FSH stimulation (and subsequent insemination) enables to amplify the lineage of chosen females (via embryo transfer) and to preserve the genetic diversity (via embryo storage in cryobanks). The large variability in the individual responses to ovarian stimulation treatment hampers both their therapeutic and farming applications. Improving the knowledge upon the mechanisms underlying FSH control will help to improve the success of assisted reproductive technologies, hence to prevent ovarian failure or hyperstimulation syndrome in women and to manage ovulation rate and ovarian cycle chronology in farm species.

To control ovarian cycle and ovulation, we have to deeply understand the selection process of ovulatory follicles, the determinism of the species-specific ovulation rate and of its intra- and between-species variability, as well as the triggering of the ovulatory GnRH surge from hypothalamic neurons.
Beyond the strict scope of Reproductive Physiology, this understanding raises biological questions of general interest, especially in the fields of

Molecular and Cellular Biology. The granulosa cell, which is the primary target of FSH in ovarian follicles, is a remarkable cellular model to study the dynamical control of the transitions between the cellular states of quiescence, proliferation, differentiation, and apoptosis, as well as the adaptability of the response to the same extra-cellular signal according to the maturity level of the target cell. Moreover, the FSH receptor belongs to the seven transmembrane spanning receptor family, which represent the most frequent target (over 50%) amongst the therapeutic agents currently available. The study of FSH receptor-mediated signaling is thus not only susceptible to allow the identification of relaying controls to the control exerted by FSH, but it is also interesting from a more generic pharmacological viewpoint.

Neuroendocrinology and Chronobiology. The mechanisms underlying the GnRH ovulatory surge involve plasticity phenomena of both neuronal cell bodies and synaptic endings comparable to those occurring in cognitive processes. Many time scales are interlinked in ovulation control from the fastest time constants of neuronal activation (millisecond) to the circannual variations in ovarian cyclicity. The influence of daylength on ovarian activity is an interesting instance of a circannual rhythm driven by a circadian rhythm (melatonin secretion from the pineal gland).

Simulation and control of a multiscale conservation law for follicular cells

In the past years, we have designed a multiscale model of the selection process of ovulatory follicles, including the cellular, follicular and ovarian levels [11], [10]. The model results from the double structuration of the granulosa cell population according to the cell age (position within the cell cycle) and to the cell maturity (level of sensitivity towards hormonal control). In each ovarian follicle, the granulosa cell population is described by a density function whose changes are ruled by conservation laws. The multiscale structure arises from the formulation of a hierarchical control operating on the aging and maturation velocities as well on the source terms of the conservation law. The control is expressed from different momentums of the density leading to integro-differential expressions.
Future work will take place in the REGATE project and will consist in:
- predicting the selection outcome (mono-, poly-ovulation or anovulation / ovulation chronology) resulting from given combinations of parameters and corresponding to the subtle interplay between the different organs of the gonadotropic axis (hypothalamus, pituitary gland and ovaries). The systematic exploration of the situations engendered by the model calls for the improvement of the current implementation performances. The work will consist in improving the precision of the numerical scheme, in the framework of the finite volume method and to implement the improved scheme,

- solving the control problems associated with the model. Indeed, the physiological conditions for the triggering of ovulation, as well as the counting of ovulatory follicles amongst all follicles, define two nested and coupled reachability control problems. Such particularly awkward problems will first be tackled from a particle approximation of the density, in order to design appropriate control laws operating on the particles and allowing them to reach the target state sets.

Connectivity and dynamics of the FSH signaling network in granulosa cells

The project consists in analyzing the connectivity and dynamics of the FSH signaling network in the granulosa cells of ovarian follicles and embedding the network within the multiscale representation described above, from the molecular up to the organic level. We will examine the relative contributions of the Gα_s and βarrestin-dependent pathways in response to FSH signal, determine how each pathway controls downstream cascades and which mechanisms are involved in the transition between different cellular states (quiescence, proliferation, differentiation and apoptosis). On the experimental ground, we propose to develop an antibody microarray approach in order to simultaneously measure the phosphorylation levels of a large number of signaling intermediates in a single experiment. On the modeling ground, we will use the BIOCHAM (biochemical abstract machine) environment first at the boolean level, to formalize the network of interactions corresponding to the FSH-induced signaling events on the cellular scale. This network will then be enriched with kinetic information coming from experimental data, which will allow the use of the ordinary differential equation level of BIOCHAM. In order to find and fine-tune the structure of the network and the values of the kinetic parameters, model-checking techniques will permit a systematic comparison between the model behavior and the results of experiments. In the end, the cell-level model should be abstracted to a much simpler model that can be embedded into a multiscale one without losing its main characteristics.

Bifurcations in coupled neuronal oscillators.

We have proposed a mathematical model allowing for the alternating pulse and surge pattern of GnRH (Gonadotropin Releasing Hormone) secretion [5]. The model is based on the coupling between two systems running on different time scales. The faster system corresponds to the average activity of GnRH neurons, while the slower one corresponds to the average activity of regulatory neurons. The analysis of the slow/fast dynamics exhibited within and between both systems allows to explain the different patterns (slow oscillations, fast oscillations and periodical surge) of GnRH secretion.

This model will be used as a basis to understand the control exerted by ovarian steroids on GnRH secretion, in terms of amplitude, frequency and plateau length of oscillations and to discriminate a direct action (on the GnrH network) from an indirect action (on the regulatory network) of steroids. From a mathematical viewpoint, we have to fully understand the sequences of bifurcations corresponding to the different phases of GnRH secretion. This study will be derived from a 3D reduction of the original model.

STEEP Exploratory Action

3. Scientific Foundations

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 [2], 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 [3]. 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.

[2] http://www.modelistica.com/english
[3] 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. 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 cover model (CLUE-S [4]) which belongs to the last category.

In the following, our three main research axes are described.

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 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

[4] http://www.ivm.vu.nl/en/Organisation/departments/spatial-analysis-decision-support/Clue

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.

3.4. Modeling of socio-economic and environmental interactions

Considering the assessment of socio-economic impacts on the environment and ecosystem service analysis, the problems encountered here are intrinsically interdisciplinary: they draw on social sciences, ecology or Earth sciences. The modeling of the considered phenomena must take into account many factors of different nature which interact *via* various 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. The spatial processes involve proximity relationships and neighborhoods, like for example, between two adjacent parcels of land. 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. The multi-scale approaches can also be justified by the lack of data at the relevant scales. This is for example the case for the material flow analysis at local scales for which complex data disaggregations are required.

At this stage, it is crucial to understand that the scientific fields considered here are far from being mature. For example, the very notions of ecosystem services or local ecological accounting are quite recent and at best partially documented, but advances in those fields are essential, and will be required to identify transition paths to sustainability. Nowadays, the analyses are only qualitative or statistic. The phenomena are little understood.

Our goal here is then to do upstream research. It is to anticipate and to help the development of modeling tools that will be used tomorrow in these fields.

Developing 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; cellular automata, multi-agent models, system dynamics, or large systems of equations describing equilibrium models? 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? How to get models which automatically adapt to the granularity of the data and which are always numerically stable? 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?
Providing satisfying answers to these questions is a long term goal for STEEP.

VIRTUAL PLANTS Project-Team

3. Scientific Foundations

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 a component based, open-software platform for interdisciplinary research in plant modeling and simulation. This 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 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, LaBRI, Laboratory Jean Kuntzmann 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. Scientific Foundations

3.1. Scientific Foundations

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 tumour imaging for instance).
- In the field of image analysis and statistical modelling 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 labelling 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.

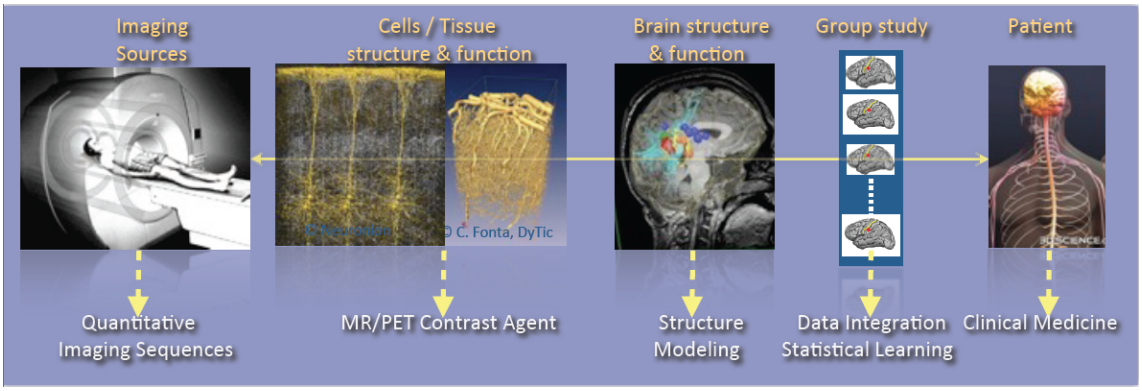


Figure 1. The major overall scientific foundation of the team concerns the integration of data from the Imaging source to the patient at different scales : from the cellular or molecular level describing the structure and function, to the functional and structural level of brain structures and regions, 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. This approaches 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.

ACES Project-Team

3. Scientific Foundations

3.1. 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. Spatial Information Systems

One of the major research efforts in ACES over the last few years has been the definition of the Spread programming model to cater for spacial context. The model is derived from the Linda [10] tuple-space model. Each information item is a *tuple*, which is a sequence of typed data items. For example, `<10, 'Peter', -3.14>` is a tuple where the first element is the integer 10, the second is the string '"Peter" and the third is the real value -3.14. Information is addressed using patterns that match one or a set of tuples present in the tuple-space. An example pattern that matches the previous tuple is `<int, 'Peter', float>`. The tuple-space model has the advantage of allowing devices that meet for the first time to exchange data since there is no notion of names or addresses.

Data items are not only addressed by their type, but also by the physical space in which they reside. The size of the space is determined by the strength of the radio signal of the device. The important difference between Spread and other tuple-space systems (e.g., Sun's JavaSpaces [9], IBM's T-Space [13]) is that when a program issues a matching request, only the tuples filling the *physical space* of the requesting program are tested for matching. Thus, though SIS (Spatial Information Systems) applications are highly distributed by nature, they only rely on localised communications; they do not require access to a global communication infrastructure. Figure 1 shows an example of a physical tuple space, made of tuples arranged in the space and occupying different spaces.

As an example of the power of this model, consider two of the applications that we have developed using it.

- *Ubi-bus* is a spatial information application whose role is to help blind and partially blind people use public transport. When taking a bus, a blind person uses his PDA to signal his intention to a device embedded in the bus stop; this device then contacts the bus on the person's behalf. This application illustrates how data is distributed over the objects of the physical world, and generally, how devices complement human means of communication.
- *Ubi-board* is a spatial information application designed for public electronic billboards. Travel hotspots like airports and major train stations have an international customer base, so bill-board announcements need to be made in several languages. In Ubi-bus, a billboard has an embedded device. When a person comes within communication range of the billboard, his device sends a request to the billboard asking it to print the message in the language of the person. In the case where several travellers are in proximity of the billboard, the board sends a translation of its information message to each person. The Ubi-board application illustrates personal context in use, i.e., the choice of natural language, and also how actions can be provoked in the physical world without explicit intervention by the person.

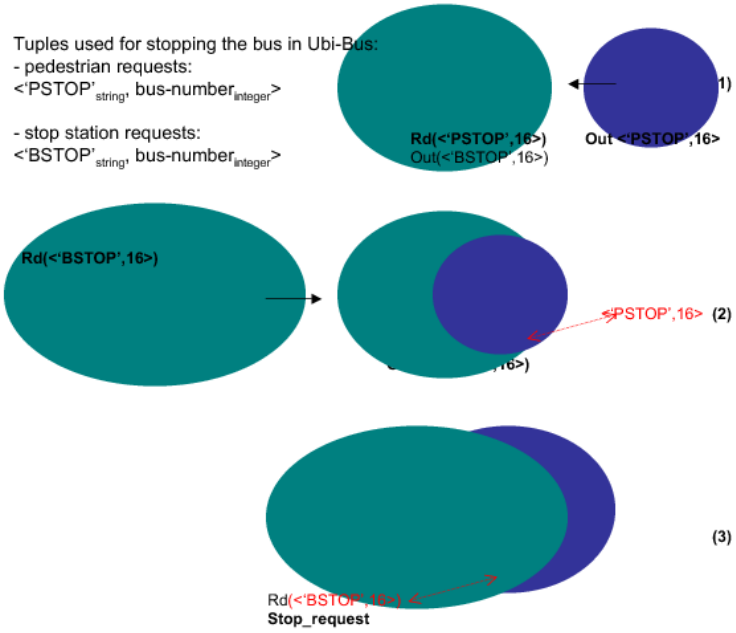


Figure 1. Physical Tuple Space

3.3. 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.

ADAM Project-Team

3. Scientific Foundations

3.1. Introduction

In order to cope with our objective, we will consider software paradigms that will help us in our approach at the various levels of our life-cycle of adaptive systems, but also in the tools themselves for their composition. We will also study these paradigms in the middleware and application design in order to extend them and to have a better understanding. These extensions will be formalized as much as possible.

3.1.1. Aspect-Oriented Software Development (AOSD)

In modern software engineering, language constructs are classified according to how they recombine partial solutions for subproblems of a problem decomposition. Some constructs (*e.g.*, methods and classes) recombine partial solutions using classic hierarchical composition. Others recombine the partial solution using what is known as crosscutting (a.k.a. aspectual) composition. With crosscutting composition, two partial solutions (called aspects) are woven into each other in a way that is dictated by so-called pointcut languages. The necessity of crosscutting composition is the main motivation for the AOSD [87], [105] paradigm. The challenge will be first to study new expressive pointcut languages in order to have a better description of composition locations in adaptable software. The second objective will be to extend and to integrate new techniques of weaving at design time, but also at run time in order to compose software safely. The third objective will be to go beyond simple aspects as persistence and logging services. We plan to study complex aspects such as transactions or replication and to control their weaving in order to master the evolution of complex software.

3.1.2. Component-Based Software Engineering (CBSE)

In a post-object world [101], software components [110] are, with other artifacts such as aspects, one of the approaches that aims at overcoming the limitations of objects and providing more flexibility and dynamicity to complex applications. For that, software components present many interesting properties, such as modularity, encapsulation, and composability. Yet, many different component models and frameworks exist. A survey of the literature references more than 20 different models (including the most well-known, such as EJB [86] and CCM [85]), but the exact number is certainly closer to 30. Indeed, each new author proposes a model to address her/his own need related to a particular execution environment (from grid computing to embedded systems) or the technical services (from advanced transactions to real-time properties), which must be provided to the application components. These different component models seldom interoperate and their design and implementation are never founded on a common ground. The research challenge that we identify is to define and implement solutions for adaptive software components. These components will be adaptive in the sense that they will be able to accommodate execution environments of various granularities (from grid computing, to Internet-based applications, to mobile applications, to embedded systems) and incorporate on-demand different technical services. This challenge will be conducted by designing a micro-kernel for software components. This micro-kernel will contain a well-defined set of core concepts, which are at the root of all component models. Several concrete software component models will then be derived from this micro-kernel.

3.1.3. Context-Aware Computing (CAC)

In adaptive systems, the notion of "*context*" becomes increasingly important. For example, mobile devices sense the environment they are in and react accordingly. This is usually enabled by a set of rules that infer how to react given a certain situation. In the Ambient/Ubiquitous/Pervasive domain [1], CAC is commonly referred to as the new paradigm that employs this idea of context in order to enmesh computing in our daily lives [113]. Many efforts that exist today focus on human-computer interaction based on context. On

[1] These terms are more or less equivalent.

the one hand, computational models, middleware, and programming languages are being developed to take the inherent characteristics of multi-scale environments into account, such as connection volatility, ambient resources, etc. An important challenge is to bridge the gap between the domain level and the computational level. The former is concerned with the expected behavior of the system from a user's viewpoint, such as how and when a system responds to changes in the context, when information can be made public, etc. On the other hand, the computational level deals with the inherent and very stringent hardware phenomena of multi-scale environments. Nevertheless, both levels have to coexist: the computational level needs to be steered by the concepts, behavior and rules which exist at the domain level, whereas the domain needs to adapt to the specificities of the ever changing environment that is monitored and managed by the computational level. In order to address this challenge, we first intend to investigate representations at the domain level of concepts such as user profile, local positioning information and execution context [126]. Furthermore, a mapping has to be devised between these concepts and generic concepts at the computational level, the latter being as independent as possible from concrete platforms or languages. This mapping has to be bidirectional: the computational level needs to be steered by the concepts, behavior and rules that exist at the domain level, whereas the domain needs to adapt to the particulars of the ever-changing environment that is monitored and managed at the computational level. Furthermore, the mapping has to be dynamic since the changes have to be propagated between the levels at run time. An explicit domain level is not only useful for bridging the aforementioned gap, but also for designing and developing open task-specific languages at the domain level, which allow users to dynamically adapt the behavior of the applications in multi-scale environments in well-defined ways.

We will base the design approach of the future implementation prototype on Model Driven Engineering (MDE). The goal of MDE [122] consists of developing, maintaining and evolving complex software systems by raising the level of abstraction from source code to models. The latter is in our case the domain level, which will be connected to the computational level by means of MDE techniques. One added benefit of MDE is that it provides means for managing model inconsistencies.

3.2. Two Research Directions

We propose to follow two research directions to foster software reuse and adaptation. The first direction, that could be coined as the spatial dimension of adaptation, will provide middleware platforms to let applications be adapted to changing execution contexts. The second direction, the so-called temporal dimension of adaptation, will provide concepts and artifacts to let designers specify evolvable applications.

3.2.1. Adaptable Component Frameworks for Middleware

As a cornerstone of next generation software, adaptation is a property which must be present throughout the entire life cycle, from design to execution. We develop then a vision where adaptation is not only a property that is desirable for end-user applications, but also for the middleware platform that executes these applications. Until now, middleware is a rather specialized activity where each new environment forces the development of a corresponding platform, which is specific to the given environment. This has led to a large number of platforms (from Web Services, to EJB, to CORBA, to ad hoc middleware for embedded systems). Although at a high level, solutions for communication interoperability often exist between these platforms, they stay loosely coupled and separated. Furthermore, the concepts which are at the core of these platforms and their architectures are too different to allow, for example, sharing technical services.

The research challenge that we propose here is to define and develop middleware and associated services which could be adapted to a broad range of environments from grid computing, to Internet-based applications, to local networks, to mobile applications on PDA's and smart phones, to embedded systems. The benefits of that are twofold. First, it enables the easier deployment of mobile applications in different environments by taking advantage of the common ground provided by adaptable middleware. Second, middleware is a rapidly changing domain where new technologies appear frequently. Yet, up to now, each new technological shift has imposed a complete re-development of the middleware. Having a common ground on which middleware is built would help in such transitions by fostering reuse. In terms of industrial output, the impact of these

results will also be helpful for software editors and companies to adapt their products more rapidly to new and emerging middleware technologies.

This research challenge has close links with MDE and product line families. We believe that the added value of our proposal is to cover a more integrated solution: we are not only interested in middleware design with MDE technologies, but we also wish to integrate them with software component technologies and advanced programming techniques, such as AOP. We will then cover a broad spectrum of middleware construction, from design (MDE) to implementation (CBSE) to application development (AOP).

3.2.2. Distributed Application Design for Adaptive Platforms

Considering adaptation in the first design steps of an application allows for its preparation and follow-up during the entire life-cycle. As mentioned previously, some software paradigms help already in the design and the development of adaptable applications. AOSD proposes separation of concerns and weaving of models in order to increase the mastering and the evolution of software. MDE consists of evolving complex software systems by raising the level of abstraction from source code to models. Several programming approaches, such as AOP or reflective approaches, have gained in popularity to implement flexibility. Other approaches, such as CBSE, propose compositional way for reuse and compose sub-systems in the application building. Finally, context-aware programming for mobile environment proposes solutions in order to consider context evolution. Overall, the objective of these approaches is to assist the development of applications that are generic and that can be adapted with respect to the properties of the domain or the context.

The research challenge that we propose to address here is similar to static points of variation in product line families. We plan to study dynamic points of variation in order to take into account adaptation in the first design steps and to match this variation. The first research challenge is the introduction of elements in the modeling phase that allow the specification of evolution related properties. These properties must make it possible to build safe and dynamic software architectures. We wish to express and validate properties in the entire software life cycle. These properties are functional, non-functional, static, behavioral, or even qualitative properties. We also want to be able to check that all the properties are present, that the obtained behavior is the expected one, and that the quality of service is not degraded after the addition or the withdrawal of functionalities. We will base our approach on the definition of contracts expressed in various formalisms (*e.g.*, first order logic, temporal logic, state automata) and we will propose a composition of these contracts.

The second challenge will be to implement design processes that maintain coherence between the various stages of modeling in a MDE approach of the applications, as well as maintaining coherence between the phases of modeling and implementation. To do so, we will design and implement tools that will enable traceability and coherence checking between models, as well as between models and the application at execution time.

Finally, we will introduce context information in the development process. At the modeling level, we will represent concepts, behavior and rules of adaptive systems to express adaptation abstraction. These models will be dynamic and connected to implementation levels at the computational level and they will consider context knowledge. The goal is to bridge the gap between the computational level and the domain level in adaptive systems by synchronization of models and implementations, but also by representation of such common knowledge.

ALGORILLE Project-Team

3. Scientific Foundations

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 (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 adaption strategies between codes and platforms: e.g the best cache strategies can't help applications that accesses 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.

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. For the architectures we will have to take into account the fact of increased heterogeneity, processors of different speed, 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 applications.

3.1.4. Frequent IO

A complete parallel application includes I/O of massive data, at an increasing frequency. In addition to applicative input and output data flow, I/O is 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 has to be included in the design of parallel programming models and tools. ORWL will 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 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 its application domain, especially concerning the detection of global convergence and the control of asynchronism. We are currently working on the design of a generic and non-intrusive way of implementing such a procedure in any parallel iterative algorithm.

Also, we would like to compare other variants of asynchronous algorithms, such as waveform relaxations. Here, computations are not performed for each time step of the simulation but for an entire time interval. Then, the evolution of the elements at the frontiers between the domain that are associated to the processors are exchanged asynchronously. Although we have already studied such schemes in the past, we would like to see how they will behave on recent architectures, and how the models and software for data consistency mentioned above can be helpful in that context.

3.1.6. Cost models and accelerators

We have already designed some models that relate computation power and energy consumption. Our next goal is to design and implement 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 able to adapt its behavior to the requirements.

3.2. Transparent Resource Management for Clouds

During the next years, we will continue to design resource provisioning strategies for cloud clients. Given the extremely large offer of resources by public or private clouds, users need software assistance to make provisioning decisions. Our goal is to gather our strategies into a **cloud resource broker** which will handle the workload of a user or of a community of users as a multi-criteria optimization problems. The notions of resource usage, scheduling, provisioning and task management have to be 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: the new prevalent optimization criterion is the financial cost of the computation.

3.2.1. Provisioning strategies

Future work includes the design of new strategies to reuse already leased resources, or switch to less powerful and cheaper resources. On one hand, some economic models proposed by cloud providers may involve a complex cost-benefit analysis for the client which we want to address. On the other hand, these economic models incur additional costs, e.g for data storage or transfer, which must 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 user workload analysis. Characterizing the workload may help to anticipate the resource provisioning, and hence improve the scheduling.

3.2.3. Experimentations

Given the very large consumption of CPU hours, the above strategies will first be tested mostly through simulation. Therefore, we will closely work with the members of the Experimental methodologies axis to co-design the cloud interface and the underlying models. Furthermore, we will assess the gap between the performances on simulation and both public and private cloud. This work will take place inside the Cloud work package of the SONGS ANR project.

3.2.4. HPC on clouds

Clouds are not suitable to run massive HPC applications. However, it might be interesting to use them as cheap HPC platform for occasional or one shot executions. This will be investigated with the Structuring Applications axis and in collaboration with the LabEx IRMIA and the CALVI team.

3.3. Experimental methodologies for the evaluation of distributed systems

We strive at designing a comprehensive set of solutions for experimentation on distributed systems by working on several methodologies (simulation, direct execution on experimental facilities, emulation) and by leveraging the convergence opportunities between methodologies (shared interfaces, validation combining several methodologies).

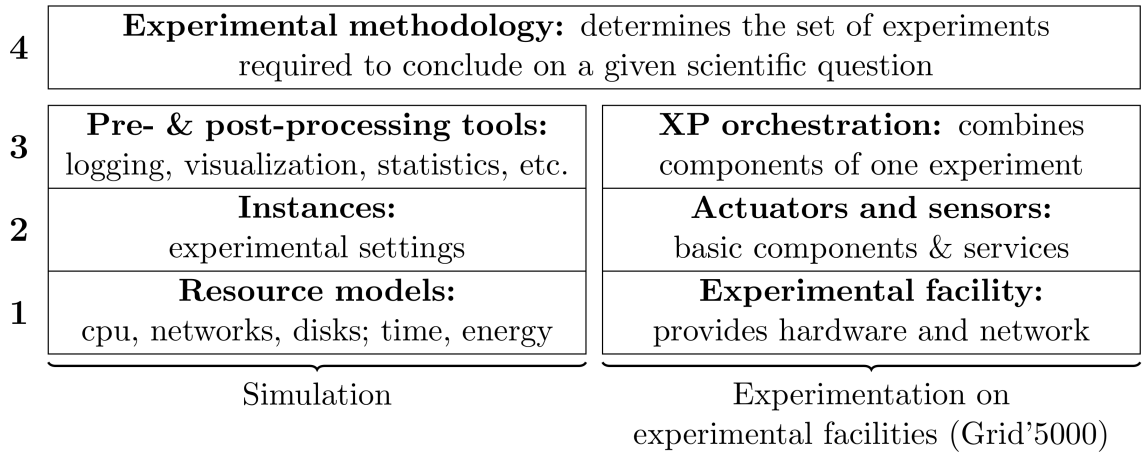


Figure 2. Our experimentation methodology, encompassing both simulation and experimental facilities.

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. Recent, we increased its audience to target the P2P research community in addition to the one on grid scheduling. It now allows **precise simulations of millions of nodes** using a single computer.

In the future, we aim at extending further the applicability to **Clouds and Exascale systems**. Therefore, we work toward disk and memory models in addition to the already 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 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 the ones developed in our team to structure applications, in particular ORWL. In the near future, we also consider using our toolbox to give **online performance predictions to the runtimes**. It would allow these systems to improve their adaptability to the changing performance conditions experienced on the platform.

We recently integrated a model checking kernel in our tool 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 using direct execution on testbeds and production facilities.

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, see Figure 3 ,

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, ...and investigate if altering memory and disk performance is possible. Other goals are to scale the tool up to 20000 virtual nodes and to improve its 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. Our approach is based on the reuse of lessons learned in the field of Business Process Management (BPM), with the design of a workflow-based experiment control engine. This is part of an ongoing collaboration with EPI SCORE (INRIA Nancy Grand Est), which has already yield promising preliminary results [15], [28].

3.3.3. Exploring new scientific objects.

We aim at addressing different kinds of distributed systems (HPC, Cloud, P2P, Grid) using the same experimental approaches. Thus a key requirement for our success is to build sufficient knowledge on target distributed systems to discover and understand the final research questions that our solutions should target. In the framework of ANR SONGS (2012-2016), we are working closely with experts from HPC, Cloud, P2P and Grid. We are also collaborating with the production grids community, e.g. on using Grid'5000 to evaluate the gLite middleware, and with Cloud experts in the context of the OpenCloudWare project.

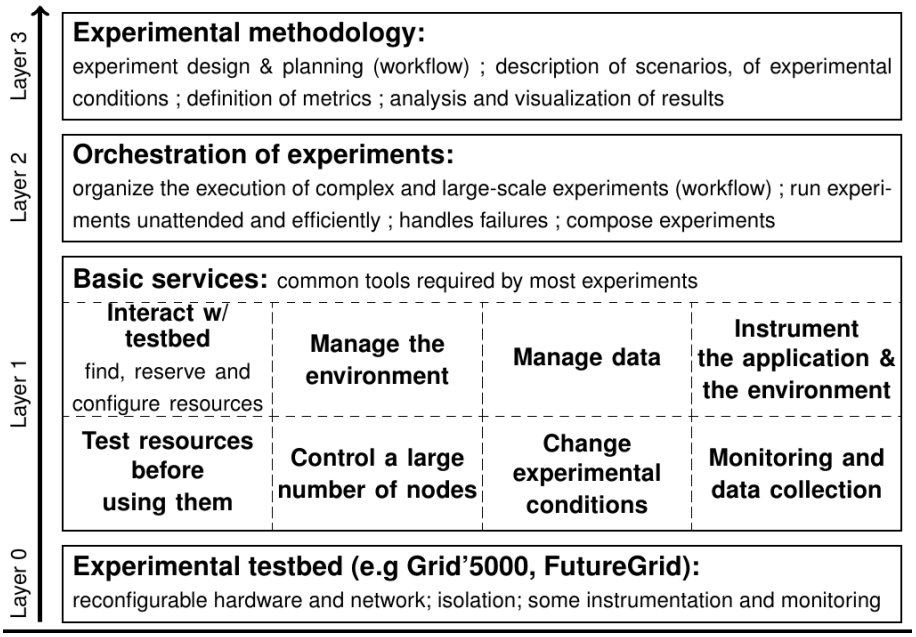


Figure 3. General structure of our project: We plan to address all layers of the experimentation stack.

ARLES Project-Team

3. Scientific Foundations

3.1. Introduction

Research undertaken within the ARLES project-team aims to offer comprehensive solutions to support the development of pervasive computing systems that are dynamically composed according to networked resources in the environment. This leads us to investigate methods and tools supporting the engineering of pervasive software systems, with a special emphasis on associated middleware solutions.

3.2. Engineering Pervasive Software Systems

Since its emergence, middleware has proved successful in assisting distributed software development, making development faster and easier, and significantly promoting software reuse while overcoming the heterogeneity of the distributed infrastructure. As a result, middleware-based software engineering is central to the principled development of pervasive computing systems. In this section, we (i) discuss challenges that middleware brings to software engineering, and (ii) outline a revolutionary approach to middleware-based software engineering aiming at the dynamic runtime synthesis of emergent middleware.

3.2.1. Middleware-based Software Engineering

Middleware establishes a new software layer that homogenizes the infrastructure's diversities by means of a well-defined and structured distributed programming model, relieving software developers from low-level implementation details, by: (i) at least abstracting transport layer network programming via high-level network abstractions matching the application computational model, and (ii) possibly managing networked resources to offer quality of service guarantees and/or domain specific functionalities, through reusable middleware-level services. More specifically, middleware defines:

- A resource definition language that is used for specifying data types and interfaces of networked software resources;
- A high-level addressing scheme based on the underlying network addressing scheme for locating resources;
- Interaction paradigms and semantics for achieving coordination;
- A transport/session protocol for achieving communication; and
- A naming/discovery protocol with related registry structure and matching relation for publishing and discovering the resources available in the given network.

Attractive features of middleware have made it a powerful tool in the software system development practice. Hence, middleware is a key factor that has been and needs to be further taken into account in the Software Engineering (SE) discipline [5]. The advent of middleware standards have further contributed to the systematic adoption of this paradigm for distributed software development.

[5] W. Emmerich. Software Engineering and Middleware: a roadmap. In Proceedings of the Conference on the Future of Software Engineering, Limerick, Ireland, Jun. 2000.

In spite of the above, mature engineering methodologies to comprehensively assist the development of middleware-based software systems, from requirements analysis to deployment and maintenance, are lagging behind. Indeed, systematic software development accounting for middleware support is rather the exception than the norm, and methods and related tools are dearly required for middleware-based software engineering. This need becomes even more demanding if we consider the diversity and scale of today's networking environments and application domains, which makes middleware and its association with applications highly complex [5], raising new, challenging requirements for middleware. Among those, access to computational resources should be open across network boundaries and dynamic due to the potential mobility of host- and user-nodes. This urges middleware to support methods and mechanisms for description, dynamic discovery and association, late binding, and loose coordination of resources. In such variable and unpredictable environments, operating not only according to explicit system inputs but also according to the context of system operation becomes of major importance, which should be enabled by the middleware. Additionally, the networking infrastructure is continuing to evolve at a fast pace, and suggesting new development paradigms for distributed systems, calling for next-generation middleware platforms and novel software engineering processes integrating middleware features in all phases of the software development.

3.2.2. Beyond Middleware-based Architectures for Interoperability

As discussed above, middleware stands as the conceptual paradigm to effectively network together heterogeneous systems, specifically providing upper layer interoperability. That said, middleware is yet another technological block, which creates islands of networked systems.

Interoperable middleware has been introduced to overcome middleware heterogeneity. However, solutions remain rather static, requiring either use of a proprietary interface or a priori implementation of protocol translators. In general, interoperability solutions solve protocol mismatch among middleware at syntactic level, which is too restrictive. This is even truer when one considers the many dimensions of heterogeneity, including software, hardware and networks, which are currently present in ubiquitous networking environments, and that require fine tuning of the middleware according to the specific capacities embedded within the interacting parties. Thus, interoperable middleware can at best solve protocol mismatches arising among middleware aimed at a specific domain. Indeed, it is not possible to a priori design a universal middleware solution that will enable effective networking of digital systems, while spanning the many dimensions of heterogeneity currently present in networked environments and further expected to increase dramatically in the future.

A revolutionary approach to the seamless networking of digital systems is to synthesize connectors on the fly, via which networked systems communicate. The resulting emergent connectors then compose and further adapt the interaction protocols run by the connected systems, from the application layer down to the middleware layer. Hence, thanks to results in this new area, networked digital systems will survive the obsolescence of interaction protocols and further emergence of new ones.

We have specifically undertaken cooperative research on the dynamic synthesis of emergent connectors which shall rely on a formal foundation for connectors that allows learning, reasoning about, and adapting the interaction behavior of networked systems [6]. Further, compared to the state of the art foundations for connectors, it should operate a drastic shift by learning, reasoning about, and synthesizing connector behavior at run-time. Indeed, the use of connector specifications pioneered by the software architecture research field has mainly been considered as a design-time concern, for which automated reasoning is now getting practical even if limitations remain. On the other hand, recent effort in the semantic Web domain brings ontology-based semantic knowledge and reasoning at run-time; however, networked system solutions based thereupon are currently mainly focused on the functional behavior of networked systems, with few attempts to capture their interaction behavior as well as non-functional properties. In this new approach, the interaction protocols (both application- and middleware-layer) behavior will be learnt by observing the interactions of the networked

[6] Valérie Issarny, Bernhard Steffen, Bengt Jonsson, Gordon S. Blair, Paul Grace, Marta Z. Kwiatkowska, Radu Calinescu, Paola Inverardi, Massimo Tivoli, Antonia Bertolino, Antonino Sabetta: CONNECT Challenges: Towards Emergent Connectors for Eternal Networked Systems. In Proceedings of ICECCS 2009.

systems, where ontology-based specification and other semantic knowledge will be exploited for generating connectors on the fly.

3.3. Middleware Architectures for Pervasive Computing

Today's wireless networks enable dynamically setting up temporary networks among mobile nodes for the realization of some distributed function. However, this requires adequate development support and, in particular, supporting middleware platforms for alleviating the complexity associated with the management of dynamic networks composed of highly heterogeneous nodes. In this section, we present an overview of: (i) service oriented middleware, a prominent paradigm in large distributed systems today, and (ii) middleware for wireless sensor networks, which have recently emerged as a promising platform.

3.3.1. Service Oriented Middleware

The *Service Oriented Computing*(SOC) paradigm advocates that networked resources should be abstracted as services, thus allowing their open and dynamic discovery, access and composition, and hence reuse. Due to this flexibility, SOC has proven to be a key enabler for pervasive computing. Moreover, SOC enables integrating pervasive environments into broader service oriented settings: the current and especially the *Future Internet* is the ultimate case of such integration. We, more particularly, envision the Future Internet as a ubiquitous setting where services representing resources, people and things can be freely and dynamically composed in a decentralized fashion, which is designated by the notion of service choreography in the SOC idiom. In the following, we discuss the role that *service oriented middleware* is aimed to have within our above sketched vision of the Future Internet, of which pervasive computing forms an integral part.

From service oriented computing to service oriented middleware: In the last few years, there is a growing interest in choreography as a key concept in forming complex service-oriented systems. Choreography is put forward as a generic abstraction of any possible collaboration among multiple services, and integrates previously established views on service composition, among which service orchestration. Several different approaches to choreography modeling can be found in the literature: *Interaction-oriented* models describe choreography as a set of interactions between participants; while *process-oriented* models describe choreography as a parallel composition of the participants' business processes. *Activity-based* models focus on the interactions between the parties and their ordering, whereas the state of the interaction is not explicitly modeled or only partly modeled using variables; while *state-based* models model the states of the choreography as first-class entities, and the interactions as transitions between states.

The above modeling categorizations are applied in the ways in which: service choreographies are specified (e.g., by employing languages such as BPMN, WS-CDL, BPEL); services are discovered, selected and composed into choreographies (e.g., based on their features concerning interfaces, behavior, and non-functional properties such as QoS and context); heterogeneity between choreographed services is resolved via adaptation (e.g., in terms of service features and also underlying communication protocols); choreographies are deployed and enacted (e.g., in terms of deployment styles and execution engines); and choreographies are maintained/adapted given the independent evolution of choreographed services (e.g., in terms of availability and QoS). These are demanding functionalities that service oriented middleware should provide for supporting service choreographies. In providing these functionalities in the context of the Future Internet, service oriented middleware is further challenged by two key Future Internet properties: its *ultra large scale* as in number of users and services, and the *high degree of heterogeneity* of services, whose hosting platforms may range from that of resource-rich, fixed hosts to wireless, resource-constrained devices. These two properties call for considerable advances to the state of the art of the SOC paradigm.

Our work in the last years has focused on providing solutions to the above identified challenges, more particularly in the domain of pervasive computing. Given the prevalence of mobile networking environments and powerful hand-held consumer devices, we consider resource constrained devices (and things, although we focus on smart, i.e., computation-enabled, things) as first-class entities of the Future Internet. Concerning middleware that enables networking mobile and/or resource constrained devices in pervasive computing environments, several promising solutions have been proposed, such as mobile Gaia, TOTA, AlfredO, or work

at UCL, Carnegie Mellon University, and the University of Texas at Arlington. They address issues such as resource discovery, resource access, adaptation, context awareness as in location sensitivity, and pro-activeness in a seamless manner. Other solutions specialize in sensor networks; we, more specifically, discuss middleware for wireless sensor networks in the next section. In this very active domain of service-oriented middleware for pervasive computing environments, we have extensive expertise that ranges from lower-level cross-layer networking to higher-level semantics of services, as well as transversal concerns such as context and privacy. We have in particular worked on aspects including semantic discovery and composition of services based on their functional properties, heterogeneity of service discovery protocols, and heterogeneity of network interfaces. Based on our accumulated experience, we are currently focusing on some of the still unsolved challenges identified above.

QoS-aware service composition: With regard to service composition in pervasive environments, taking into account QoS besides functional properties ensures a satisfactory experience to the end user. We focus here on the orchestration-driven case, where service composition is performed to fulfill a task requested by the user along with certain QoS constraints. Assuming the availability of multiple resources in service environments, a large number of services can be found for realizing every sub-task part of a complex task. A specific issue emerges in this regard, which is about selecting the best set of services (i.e., in terms of QoS) to participate in the composition, meeting user's global QoS requirements. QoS-aware composition becomes even more challenging when it is considered in the context of dynamic service environments characterized by changing conditions. As dynamic environments call for fulfilling user requests on the fly (i.e., at run-time) and as services' availability cannot be known a priori, service selection and composition must be performed at runtime. Hence, the execution time of service selection algorithms is heavily constrained, whereas the computational complexity of this problem is NP-hard.

Coordination of heterogeneous distributed systems: Another aspect that we consider important in service composition is enabling integration of services that employ different interaction paradigms. Diversity and ultra large scale of the Future Internet have a direct impact on coordination among interacting entities. Our choice of choreography as global coordination style among services should further be underpinned by support for and interoperability between heterogeneous interaction paradigms, such as message-driven, event-driven and data-driven ones. Different interaction paradigms apply to different needs: for instance, asynchronous, event-based publish/subscribe is more appropriate for highly dynamic environments with frequent disconnections of involved entities. Enabling interoperability between such paradigms is imperative in the extremely heterogeneous Future Internet integrating services, people and things. Interoperability efforts are traditionally based on, e.g., bridging communication protocols, where the dominant position is held by ESBs, wrapping systems behind standard technology interfaces, and/or providing common API abstractions. However, such efforts mostly concern a single interaction paradigm and thus do not or only poorly address cross-paradigm interoperability. Efforts combining diverse interaction paradigms include: implementing the LIME tuple space middleware on top of a publish/subscribe substrate; enabling Web services/SOAP-based interactions over a tuple space binding; and providing ESB implementations based on the tuple space paradigm.

Evolution of service oriented applications: A third issue we are interested in concerns the maintenance of service-oriented applications despite the evolution of employed services. Services are autonomous systems that have been developed independently from each other. Moreover, dynamics of pervasive environments and the Future Internet result in services evolving independently; a service may be deployed, or un-deployed at anytime; its implementation, along with its interface may change without prior notification. In addition, there are many evolving services that offer the same functionality via different interfaces and with varying quality characteristics (e.g., performance, availability, reliability). The overall maintenance process amounts to replacing a service that no longer satisfies the requirements of the employing application with a substitute service that offers the same or a similar functionality. The goal of seamless service substitution is to relate the substitute service with the original service via concrete mappings between their operations, their inputs and outputs. Based on such mappings, it is possible to develop/generate an adapter that allows the employing application to access the substitute service without any modification in its implementation. The service substitution should be dynamic and efficient, supported by a high level of automation. The state

of the art in service substitution comprises various approaches. There exist efforts, which assume that the mappings between the original and the substitute service are given, specified by the application or the service providers. The human effort required makes these approaches impractical, especially in the case of pervasive environments. On the other hand, there exist automated solutions, proposing mechanisms for the derivation of mappings. The complexity of these approaches scales up with the cardinality of available services and therefore efficiency is compromised. Again, this is an important disadvantage, especially considering the case of pervasive environments.

3.3.2. Middleware for Wireless Sensor Networks

Wireless sensor networks (WSNs) enable low cost, dense monitoring of the physical environment through collaborative computation and communication in a network of autonomous sensor nodes, and are an area of active research. Owing to the work done on system-level functionalities such as energy-efficient medium access and data-propagation techniques, sensor networks are being deployed in the real world, with an accompanied increase in network sizes, amount of data handled, and the variety of applications. The early networked sensor systems were programmed by the scientists who designed their hardware, much like the early computers. However, the intended developer of sensor network applications is not the computer scientist, but the designer of the system *using* the sensor networks, which might be deployed in a building or a highway. We use the term *domain expert* to mean the class of individuals most likely to use WSNs – people who may have basic programming skills but lack the training required to program distributed systems. Examples of domain experts include architects, civil and environmental engineers, traffic system engineers, medical system designers etc. We believe that the wide acceptance of networked sensing is dependent on the ease-of-use experienced by the domain expert in developing applications on such systems.

The obvious solution to enable this ease-of-use in application development is sensor network middleware, along with related programming abstractions [7]. Recent efforts in standardizing network-layer protocols for embedded devices provide a sound foundation for research and development of middleware that assist the sensor network developers in various aspects that are of interest to us, including the following.

Data-oriented operations: A large number of WSN applications are concerned with sampling and collection of data, and this has led to a large body of work to provide middleware support to the programmer of WSNs for easy access to the data generated and needed by the constituent nodes. Initial work included Hood, and TeenyLIME, which allowed data-sharing over a limited spatial range. Further work proposed the use of the DART runtime environment, which exposes the sensor network as a distributed data-store, addressable by using logical addresses such as “all nodes with temperature sensors in Room 503”, or “all fire sprinklers in the fifth and sixth floors”, which are more intuitive than, say, IP addresses. Taking a different approach toward handling the data in the sensor network, some middleware solutions propose to manipulate them using semantic techniques, such as in the Triple Space Computing approach, which models the data shared by the nodes in the system as RDF triples (subject-predicate-object groups), a standard method for semantic data representation. They propose to make these triples available to the participating nodes using a tuple space, thus giving it the “triple space” moniker. S-APL or Semantic-Agent Programming Language uses semantic technologies to integrate the semantic descriptions of the domain resources with the semantic prescription of agent behavior.

Integration with non-WSN nodes: Most of the work above focuses on designing applications that exhibit only intra-network interactions, where the interaction with the outside world is only in the form of sensing it, or controlling it by actuation. The act of connecting this data to other systems outside the sensor network is mostly done using an external gateway. This is then supported by middlewares that expose the sensor network as a database (e.g., TinyDB and Cougar), allowing the operator to access the data using a SQL-like syntax, augmented with keywords that can be used to specify the rate of sampling, for example. Another direction of integrating WSNs in general with larger systems such as Web servers has been toward using REST (REpresentational State Transfer) technologies, which are already used for accessing services on the Web as a

[7] L. Mottola and G. P. Picco. Programming Wireless Sensor Networks: Fundamental Concepts and State of the Art. In ACM Computing Surveys. Volume 43, Issue 3. April 2011.

lightweight alternative to SOAP. There has also been work proposing a system that will enable heterogeneous sensor and actuators to expose their sensing and actuation capabilities in a plug and play fashion. It proposes a middleware that defines a set of constraints, support services and interaction patterns that follow the REST architectural style principles, using the ATOM Web publishing protocol for service description, and a two-step discovery process. Additionally, there has been work in implementation of a REST-oriented middleware that runs on embedded devices such as Sun SPOT nodes, and the Plogg wireless energy monitors. This involves a two-fold approach — embedding tiny Web servers in devices that can host them, and employing a proxy server in situations where that is not the case. However, it has been noticed that the abstractions provided by REST might be too simplistic to compose complex applications over the services provided by WSN nodes. Some of the most recent work in this area also proposes to convert existing (network-layer) gateways into smart gateways, by running application code on them.

In addition to supporting the above interactions, sensor network middleware has also been proposed to address the challenges arising from the fact that a particular sensor or actuator may not be always available. This leads to the need for transparent reconfiguration, where the application developer should not have to care about reliability issues. The PIRATES event-based middleware for resource-rich nodes (hosting sensors/actuators, or just processing data) includes a third-party-remapping facility that can be used to remap a component's endpoints without affecting the business logic. In that sense, it is similar to the RUNES middleware targeted at embedded systems.

Finally, we also note the recent initial WSN middleware research focused on the new nascent classes of systems. Most recently, the field of *participatory sensing*[8] has emerged, where the role of sensing is increasingly being performed by the mobile phones carried by the users of the system, providing data captured using the sound, GPS, accelerometer and other sensors attached to them. This has led to the emergence of middleware such as JigSaw. The core additional challenges in this domain come from the inherent mobility of the nodes, as well as their extremely large scale.

[8]Lane, N.D.; Miluzzo, E.; Hong Lu; Peebles, D.; Choudhury, T.; Campbell, A.T.; , "A survey of mobile phone sensing," Communications Magazine, IEEE , vol.48, no.9, pp.140-150, Sept. 2010

ASAP Project-Team

3. Scientific Foundations

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*" [56] (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 [57], [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 thousand 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 [58], [59], [60] following a specific topology or an unstructured network [55], [61] 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 the 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 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

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.

ASCOLA Project-Team

3. Scientific Foundations

3.1. Overview

Since we mainly work on new software structuring concepts and programming language design, we first briefly introduce some basic notions and problems of software components (understood in a broad sense, i.e., 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 that are relevant to our work.

3.2. Software Components

Modules and services. The idea that building *software components*, i.e., composable prefabricated and parametrized software parts, was key to create an effective software industry was realized very early [67]. 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 [76], [48]. 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 a first step, it is possible to reason locally, about the consistency between the module implementation and the module interface. In a second step, it is possible to reason about composing modules by only considering their interfaces. Modern module systems also consider types as module elements and consider, typically static, modules as a unit of separate compilation, with the most sophisticated ones also supporting modules parametrized by modules [65].

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 [71].

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* [79] of a software system describes the system as a composition of *components* and *connectors*, where the connectors capture the *interaction protocols* between the components [39]. 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 [68]. 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.

3.3. Aspect-Oriented Programming

The main driving force for the structuring means, such as components and modules, is the quest for clean *separation of concerns* [50] 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 [80], [62] 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.

Aspect-Oriented Software Development [61], [37] 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 [60], 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 [81]. 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 [47]. 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 [83]) and provide only limited support for the analysis of fundamental aspect properties.

Recently, first approaches have been put forward to tackle these problems, in particular, in the context of so-called *stateful* or *history-based aspect languages* [51], [52], 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 [75], [41], [54], provide more concise formal semantics for aspects and enable analysis of their properties [40], [53], [51], [38]. Due to the novelty of these approaches, they represent, however, only first results and many important questions concerning these fundamental issues remain open.

3.4. Protocols

Today, protocols constitute a frequently used means to precisely define, implement, and analyze contracts 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 [74], component interactions [84], [77] and service orchestrations [49] 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 [44], [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 [59], or context-free or turing-complete expressiveness [78], [43]. 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.

Furthermore, first approaches for the definition of *aspects over protocols* have been proposed, as well as over regular structures [51] and non-regular ones [82], [73]. The modification of interaction protocols by aspects seems highly promising for the *integration of aspects and components*.

3.5. Patterns

Patterns provide a kind of abstraction that is complementary to the modularization mechanisms discussed above. They have been used, in particular, to define general *architectural styles* either by defining entire computation and communication topologies [72], connectors between (complex) software artifacts [69], or (based on, possibly concretizations of, *design patterns* [58]) as building blocks for object-oriented software architectures. The resulting pattern-based architectures are similar to common component-based architectures and are frequently used to implement the latter, see, for instance, Sun's J2EE patterns.

Patterns have also been used to implement architectural abstractions. This is the case, for instance, for the numerous variants of the *publish/subscribe pattern* [55] as well as the large set of so-called *skeletons* [46], that is, patterns for the implementation of distributed and concurrent systems. While these patterns are essentially similar to architecture-level patterns, their fine-grained application to multiple code entities often results in crosscutting code structures. An important open issue consists in the lack of pattern-based representations for the implementation of general distributed applications — in sharp contrast to their use for the derivation of massively parallel programs.

3.6. Domain-Specific Languages

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 [70], [85].

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.

3.7. 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* [42], [64]. 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 [63] and the modularization of the distribution concern itself [81]. 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 [40], [54].

ATLANMOD Team

3. Scientific Foundations

3.1. MDE Foundations

MDE can be seen as a generalization and abstraction of object technology allowing to map more abstract organizations on class-based implementations. In MDE, (software) models are considered as the unifying concept [44].

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. 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 [52] and [45]. MDE promotes *unification by models*, like object technology proposed in the eighties *unification by objects* [43]. 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 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 [47]. 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 [4]. 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 [46]. 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 [8]).

When following the previously described principles, one may take advantage of the uniformity of the MDE organization. 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. Considering models of products (e.g., software artifacts like UML) and models of processes (e.g., software processes like SPEM) may lead to a dual process/product organization. Considering transformation models, weaving models, and traceability models as special cases of correspondence models may also lead to simplicity and efficiency of implementations. These are some of the use cases that are being explored in the team.

AVALON Team

3. Scientific Foundations

3.1. Algorithmics

The researches conducted by the Avalon team address both complex applications, coming from service/component composition and more generally organized as workflows, and complex architectures, that are heterogeneous, distributed, shared, and elastic. While some characteristics are classical to parallel and distributed platforms such as Clusters and Grids, new challenges arise because of the increase of complexity of application structures as well as by the elasticity of infrastructures such as Clouds and by the importance of taking into account energy concerns in Supercomputers for example.

Moreover data-intensive applications imply not only to consider computations in a scheduling process but also data movements in a coordinated way.

In such a context, many metrics can be optimized by transformation and/or scheduling algorithms in order to deploy services or applications on resources. Classical ones are the minimization of application completion or turnaround times, the maximization of the resource usage, or taking care of the fairness between applications. But new challenging optimizations are now related to the economical cost of an execution or to its energy efficiency.

Our main challenge is to propose smart transformation and scheduling algorithms that are inherently multi-criteria optimizations. As not all metrics can be simultaneously optimized, the proposed algorithms consider subsets of them: we target at finding efficient trade-offs. Note that our main concern is to design practical algorithms rather than conducting purely theoretical studies as our goal is at implementing the proposed algorithms in actual software environments.

Moreover, in recent years, we have seen the apparition of hardware-based green leverages (on/off, idle modes, dynamic frequency and speed scaling, etc) applied to various kinds of physical resources (CPU, memory, storage and network interconnect). To exploit them, these facilities must be incorporated into middleware software layers (schedulers, resource managers, etc). The Avalon team explores the benefits of such leverages, for example with respect to elasticity, to improve the energy efficiency of distributed applications and services and to limit the energy consumption of platforms. The goal is to provide the needed amount of physical and virtual resources to fulfill the needs of applications. Such provision is greatly influenced by a large set of contextual choices (hardware infrastructures, software, location, etc).

3.2. Application and Resource Models

A second research direction consists in providing accurate, or at least realistic, models of applications and execution infrastructures. Such a goal has been the main concern of the *SimGrid* project for more than 10 years. Hence, this simulation toolkit provides most of the technological background to allow for the exploration of new scientific challenges. Moreover, simulation is a classical and efficient way to explore many “what-if” scenarios in order to better understand how an application behaves under various experimental conditions.

The Avalon team considers using simulation for application performance prediction. The scientific challenges lie in the diversity of applications and available execution environments. Moreover the behavior and performance of a given application may vary greatly if the execution context changes. Simulation allows us to explore many scenarios in a reasonable time, but this require to get a good understanding of both application structure and target environment.

A first focus is on HPC, regular, and parallel applications. For instance, we study those based on the message passing paradigm, as we have already developed some online and offline simulators. However, the different APIs provided by *SimGrid* allow us to also consider other kinds of applications, such as scientific workflows or CSPs.

A second focus is on data-intensive applications. It implies to also consider storage elements as a main modeling target. In the literature, the modeling of disk is either simplistic or done at a very-low level. This leads to unrealistic or intractable models that prevent the acquisition of sound information. Our goal is then to propose comprehensive models at the storage system level, *e.g.,* one big disk bay accessed through the network. The main challenge associated to this objective is to analyze lots of logs of accesses to data to find patterns and derive sound models. The IN2P3 Computing Center gives us an easy access to such logs. Moreover a collaboration with CERN will allow us to validate the proposed model on an actual use case, the distributed data management system of the ATLAS experiment.

Modeling applications and infrastructures is in particular required to deal with energy concerns, as energy price is becoming a major limiting factor for large scale infrastructures. Physically monitoring the energy consumption of few resources is now becoming a reality; injecting such local measurements as a new parameter in multi-objective optimization models is also more and more common. However, dealing with energy consumption and energy efficiency at large scale is still a real challenge. This activity, initiated in the RESO team since 2008, is continued by the Avalon team by investigating energy consumption and efficiency on large scale (external, internal) monitoring of resources. Also, while physical resources start to be well mastered, another challenge is to deal with virtualized resources and environments.

3.3. Programming Abstractions

Another research direction deals with determining well suited programming abstractions to reconcile a priori contradictory goals: being “simple” to use and portable, while enabling high performance. Existing parallel and distributed programming models either expose infrastructure artifacts to programmers so that performance can be achieved —by experts!— but not portability or they propose very specialized models such as GridRPC and Google’s MapReduce. In the latter case, an application is restricted to use one concept at a time. For example, it is very difficult for an application to simultaneously use two middleware systems providing respectively GridRPC and MapReduce.

The Avalon team addresses the challenge of designing a general composition based model supporting as many composition operators as possible while enabling efficient execution on parallel and distributed infrastructures. We mainly consider component based models 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 the best suited models.

Many works have already been done with respect to component models. However, existing model such as Fractal, CCA, BIP do not provide an adequate solution as they only support a limited set of interactions. We aims to extend the approach initiated with HLCM that aims at identifying core elements to let a programmer define any (spatial) compositions. We also target to provide mechanisms to support application and resource specialization algorithms.

A first challenge is to conduct an in depth validation of the ability of the proposed HLCM approach to deal with any kind of static compositions. In particular, it includes designing efficient transformation algorithms and understanding their generality.

A second challenge is to extend the proposed approach to support dynamic applications, either because of adaptation issues or because of temporal compositions such as workflows. Starting from motivating use cases, we will study whether just-in-time assembly transformation techniques provide an efficient and scalable solution.

3.4. Resource abstractions

Computing resources and infrastructures have a wide variety of characteristics in terms of reliability, performance, service quality, price, energy consumption, etc. Moreover, resource usage and access differ from batch scheduler, reservation, on-demand, best-effort, virtualized, etc.

The Avalon team addresses issues related to the provision of the necessary resource abstractions to allow efficient resource usage, the accurate description of resource properties, and the efficient management of the complexity of hybrid distributed infrastructures.

The challenge is threefold: $i)$ providing the adequate resource management services to cope with large scale, heterogeneous, volatile, and elastic infrastructures, $ii)$ combining several DCIs together, $iii)$ providing feedback on how applications make use of resources, which implies for instance energy monitoring facility.

The Avalon team aims at designing and evaluating adapted services such as job scheduler, decentralized resource discovery, data management, monitoring systems, or QoS services. Moreover, the team studies at which level in the design stack advanced features, such as QoS, reliability, security, have to/can be provided.

Our methodology consists in designing experiments involving the investigated services. Therefore, the team closely collaborates with large-scale infrastructure operators and designers such as CC-IN2P3, GRID'5000, FutureGrid or the International Desktop Grid Federation. We aim at making use of existing DCIs services as much as possible and develop new services otherwise. In the past years, the team members have gained a recognized experience in designing middleware systems for distributed and parallel computing that rely on different resource abstractions: data management and data-intense computing (BitDew, DIET), workflows (DIET), component model (HLCM). In the next years, we plan to improve these systems or develop new services with respect to challenges related on determining how resources are found, queried, accessed, used, and released. For example, the Avalon team contributes to energy monitoring services as well as to information services, and job submission services for elastic resources.

CEPAGE Project-Team

3. Scientific Foundations

3.1. Modeling Platform Dynamics

Modeling the platform dynamics in a satisfying manner, in order to design and analyze efficient algorithms, is a major challenge. In distributed platforms, the performance of individual nodes (be they computing or communication resources) will fluctuate; in a fully dynamic platform, the set of available nodes will also change over time, and algorithms must take these changes into account if they are to be efficient.

There are basically two ways one can model such evolution: one can use a *stochastic process*, or some kind of *adversary model*.

In a stochastic model, the platform evolution is governed by some specific probability distribution. One obvious advantage of such a model is that it can be simulated and, in many well-studied cases, analyzed in detail. The two main disadvantages are that it can be hard to determine how much of the resulting algorithm performance comes from the specifics of the evolution process, and that estimating how realistic a given model is – none of the current project participants are metrology experts.

In an adversary model, it is assumed that these unpredictable changes are under the control of an adversary whose goal is to interfere with the algorithms efficiency. Major assumptions on the system's behavior can be included in the form of restrictions on what this adversary can do (like maintaining such or such level of connectivity). Such models are typically more general than stochastic models, in that many stochastic models can be seen as a probabilistic specialization of a nondeterministic model (at least for bounded time intervals, and up to negligible probabilities of adopting "forbidden" behaviors).

Since we aim at proving guaranteed performance for our algorithms, we want to concentrate on suitably restricted adversary models. The main challenge in this direction is thus to describe sets of restricted behaviors that both capture realistic situations and make it possible to prove such guarantees.

3.2. Models for Platform Topology and Parameter Estimation

On the other hand, in order to establish complexity and approximation results, we also need to rely on a precise theoretical model of the targeted platforms.

- At a lower level, several models have been proposed to describe interference between several simultaneous communications. In the 1-port model, a node cannot simultaneously send to (and/or receive from) more than one node. Most of the "steady state" scheduling results have been obtained using this model. On the other hand, some authors propose to model incoming and outgoing communication from a node using fictitious incoming and outgoing links, whose bandwidths are fixed. The main advantage of this model, although it might be slightly less accurate, is that it does not require strong synchronization and that many scheduling problems can be expressed as multi-commodity flow problems, for which efficient decentralized algorithms are known. Another important issue is to model the bandwidth actually allocated to each communication when several communications compete for the same long-distance link.
- At a higher level, proving good approximation ratios on general graphs may be too difficult, and it has been observed that actual platforms often exhibit a simple structure. For instance, many real life networks satisfy small-world properties, and it has been proved, for instance, that greedy routing protocols on small world networks achieve good performance. It is therefore of interest to prove that logical (given by the interactions between hosts) and physical platforms (given by the network links) exhibit some structure in order to derive efficient algorithms.

3.3. General Framework for Validation

3.3.1. Low level modeling of communications

In the context of large scale dynamic platforms, it is unrealistic to determine precisely the actual topology and the contention of the underlying network at application level. Indeed, existing tools such as Alnem [103] are very much based on quasi-exhaustive determination of interferences, and it takes several days to determine the actual topology of a platform made up of a few tens of nodes. Given the dynamism of the platforms we target, we need to rely on less sophisticated models, whose parameters can be evaluated at runtime.

Therefore, we propose to model each node using a small set of parameters. This is related to the theoretical notion of distance labeling [92], and corresponds to assigning labels to the nodes, so that a cheap operation on the labels of two nodes provides an estimation of the value of a given parameter (the latency or the bandwidth between two nodes, for instance). Several solutions for performance estimation on the Internet are based on this notion, under the terminology of Network Coordinate Systems. Vivaldi [83], IDES [104] and Sequoia [106] are examples of such systems for latency estimation. In the case of bandwidth estimation, fewer solutions have been proposed. We have studied the last-mile model, in which we model each node by an incoming and an outgoing bandwidth and neglect interference that appears at the core of the network (Internet), in order to concentrate on local constraints.

3.3.2. Simulation

Once low level modeling has been obtained, it is crucial to be able to test the proposed algorithms. To do this, we will first rely on simulation rather than direct experimentation. Indeed, in order to be able to compare heuristics, it is necessary to execute those heuristics on the same platform. In particular, all changes in the topology or in the resource performance should occur at the same time during the execution of the different heuristics. In order to be able to replicate the same scenario several times, we need to rely on simulations. Moreover, a metric for providing approximation results in the case of dynamic platforms necessarily requires computing the optimal solution at each time step, which can be done off-line if all traces for the different resources are stored. Using simulation rather than experiments can be justified if the simulator itself has been proven valid. Moreover, the modeling of communications, processing and their interactions may be much more complex in the simulator than in the model used to provide a theoretical approximation ratio, such as in SimGrid. In particular, sophisticated TCP models for bandwidth sharing have been implemented in SimGRID.

During the course of the USS-SimGrid ANR Arpege project, the SimGrid simulation framework has been adapted to large scale environments. Thanks to hierarchical platform description, to simpler and more scalable network models, and to the possibility to distribute the simulation of several nodes, it is now possible to perform simulations of very large platforms (of the order of 10^5 resources). This work will be continued in the ANR SONGS project, which aims at making SimGrid usable for Next Generation Systems (P2P, Grids, Clouds, HPC). In this context, simulation of exascale systems are envisioned, and we plan to develop models for platform dynamicity to allow realistic and reproducible experimentation of our algorithms.

3.3.3. Practical validation and scaling

Finally, we propose several applications that will be described in detail in Section 5. These applications cover a large set of fields (molecular dynamics, continuous integration...). All these applications will be developed and tested with an academic or industrial partner. In all these collaborations, our goal is to prove that the services that we propose can be integrated as steering tools in already developed software. Our goal is to assert the practical interest of the services we develop and then to integrate and to distribute them as a library for large scale computing.

At a lower level, in order to validate the models we propose, i.e. make sure that the predictions given by the model are close enough to the actual values, we need realistic datasets of network performance on large scale distributed platforms. Latency measurements are easiest to perform, and several datasets are available to researchers and serve as benchmarks to the community. Bandwidth datasets are more difficult to obtain, because of the measurement cost. As part of the bedibe software (see section 5.4), we have implemented a

script to perform such measurements on the Planet-Lab platform [72]. We plan to make these datasets available to the community so that they can be used as benchmarks to compare the different solutions proposed.

CIDRE Project-Team

3. Scientific Foundations

3.1. Introduction

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.* [56], 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 in a complementary manner both the concerns of accidental and intentional failures.

The second main characteristic of the CIDRE group lies in the scope of the systems it considers. Indeed, during our research, we will consider three complementary levels of study: the Node Level, the Group Level, and the Open Network Level:

- 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.
- 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.

- 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 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, i.e., trust, intrusion detection, and privacy, and on the different 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 to rely 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 Protection which is now recognized as a basic user right, should be respected despite the presence of tools 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 the preventive security mechanisms and violate the security policy of the whole system. The goal of intrusion detection systems (IDS) is to be able 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, i.e. the anomaly approach, which consists in detecting a deviation from a referenced behavior. Specifically, we propose to study two complementary methods:

- Illegal Flow Detection: This first method intends to detect information flows that violate the security policy [59], [55]. 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 [54], [58]. 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.

In both 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 first two levels. Moreover, we plan to work on intrusion detection system evaluation methods. For that research, we set a priori aside no particular IDS approach or technique. Here are some concrete examples of our research goals (both short term and long term objectives) in the intrusion detection field:

- at node level, we are going to 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 plan to extend this method to build anti-viruses and DBMS tools by determining viruses signatures.

- at the set of nodes level, we are going to 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, one of the challenges is to enforce the collaboration and consistency at nodes and set of nodes levels to obtain a global intrusion detection mechanism. Regarding the data corruption detection approach, the 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. In forthcoming years, 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, problematics 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). 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. In the CIDRE project, we will investigate PETs that operate at the three different levels (node, set of nodes or open distributed system) and 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 where privacy and utility aspects collide and that will be studied within the context of CIDRE include: identity and privacy, geo-privacy, distributed computing and privacy, privacy-preserving data mining and privacy issues in social networks. Here are some concrete examples of our research goals in the privacy field:

- at the node level, we aim at designing privacy preserving identification scheme, automated reasoning on privacy policies [57], and policy-based adaptive PETs.
- at the set of nodes level, we plan to augment distributed algorithms (i.e., consensus) with privacy properties such as anonymity, unlinkability, and unobservability.
- at the open distributed system level, we plan to target both geo-privacy concerns (that typically occur in geolocalized systems) and privacy issues in social networks. In the former case, we will adopt a sanitization approach while in the latter one we plan to define privacy policies at user level, and their enforcement by all the intervening actors (e.g, at the social network sites providers).

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) that implement 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 in order to strengthen each node's TCB. Accordingly, the CIDRE group proposes to study 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. By way of examples of our research goals regarding the trust management field, we briefly list some of our short and long term objectives at node, group and open networks levels:

1. at node level, we are going to 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 plan to 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 are going to 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 oblivion (i.e., the right to data erasure).

DANTE Team

3. Scientific Foundations

3.1. Statistical Characterization of Complex Interaction Networks

Participants: Christophe Crespelle, Éric Fleury, Adrien Friggeri, Paulo Gonçalves, Qinna Wang, Lucie Martinet, 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.

The dynamics of complex networks often exhibit no preferred time scale or equivalently involve a whole range of scales and are characterized by a scaling or scale invariance property. Another important aspect of network dynamics resides in the fact that the sensors measure information of different nature. For instance, in the MOSAR project, inter-individual contacts are registered, together with the health status of each individual, and the time evolution of the resistance to antibiotics of the various strains analyzed. Moreover, such information is collected with different and unsynchronized resolutions in both time and space. This property, referred to as multi-modality, is generic and central in most dynamical networks. With these main challenges in mind, we define the following objectives.

From "primitive" to "analyzable" data: Observables. The various and numerous modalities of information collected on the network generate a huge "*primitive*" data set. It has first to be processed to extract "*analyzable data*", which can be envisioned with different time and space resolutions: it can concern either local quantities, such as the number of contacts of each individual, pair-wise contact times and durations, or global measures, *e.g.*, the fluctuations of the average connectivity. The first research direction consists therefore in identifying, from the "primitive data", a set of "analyzable data " whose relevance and meaningfulness for the analysis of network dynamic and network diffusion phenomena will need to be assessed. Such " analyzable data " needs also to be extracted from large " primitive data " set with " reasonable " complexity, memory and computational loads.

Granularity and resolution. The corresponding data will take the form of time-series, "condensing" network dynamics description at various granularity levels, both in time and space. For instance, the existence of a contact between two individuals can be seen as a link in a network of contacts. Contact networks corresponding to contact sequences aggregated at different analysis scales (potentially ranging from hours to days or weeks) can be built. However, it is so far unclear to which extent the choice of the analysis scale impacts the relevance of network dynamics description and analysis. An interesting and open issue lies in the understanding of the evolution of the network from a set of isolated contacts (when analyzed with low resolution) to a globally interconnected ensemble of individuals (at large analysis scale). In general, this raises the question of selecting the adequate level of granularity at which the dynamics should be analyzed. This difficult problem is further complicated by the multi-modality of the data, with potentially different time resolutions.

(non-)Stationarity. Stationarity of the data is another crucial issue. Usually, stationarity is understood as a time invariance of statistical properties. This very strong definition is difficult to assess in practice. Recent efforts have put forward a more operational concept of relative stationarity in which an observation scale is explicitly included. The present research project will take advantage of such methodologies and extend them to the network dynamics context.

The rationale is to compare local and global statistical properties at a given observation scale in time, a strategy that can be adapted to the various time series that can be extracted from the data graphs so as to capture their dynamics. This approach can be given a statistical significance via a test based on a data-driven characterization of the null hypothesis of stationarity.

Dependencies, correlations and causality. To analyze and understand network dynamics, it is essential that (statistical) dependencies, correlations and causalities can be assessed among the different components of the " analyzable data ". For instance, in the MOSAR framework, it is crucial to assess the form and nature of the dependencies and causalities between the time series reflecting e.g., the evolution along time of the strain resistance to antibiotics and the fluctuations at the inter-contact level. However, the multimodal nature of the collected information together with its complex statistical properties turns this issue into a challenging task. Therefore, Task1 will also address the design of statistical tools that specifically aim at measuring dependency strengths and causality directions amongst mutivariate signals presenting these difficulties. The objective is to provide elements of answers to natural yet key questions such as : Does a given property observed on different components of the data result from a same and single network mechanism controlling the ensemble or rather stem from different and independent causes? Do correlations observed on one instance of information (e.g., topological) command correlations for other modalities? Can directionality in correlations (causality) be inferred amongst the different components of multivariate data? These should also shed complementary lights on the difficulties and issues associated to the identification of " important " nodes or links...

3.2. Theory and Structural Dynamic Properties of dynamic Networks

Participants: Christophe Crespelle, Éric Fleury, Qinna Wang, Adrien Friggeri.

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.

To go further in the Classical graph notions like the definition of path, connected components and k-core have to be revisited in this context. Advanced properties need also to be defined in order to apprehend the intrinsic dynamic structural issues of such dynamic graphs. The notion of communities (dense group of nodes) is important in any social / interaction network context and may play an important role within an epidemic process. To transpose the static graph community concept into the dynamical graph framework is a challenging task and appears necessary in order to better understand how the structure of graphs evolves in time. In these context we define the following objectives:

Toward a dynamic graph model and theory. We want to design new notions, methods and models for the analysis of dynamic graphs. For the static case, graph theory has defined a vast and consistent set of notions and methods such as paths, flows, centrality measures. These notions and methods are completely lacking for the study of dynamic graphs. We aim at providing such notions in order to study the structure of graphs evolving in time and the phenomenon taking place on these dynamic graphs. Our approach relies on describing a dynamic graph by a series of graphs which are the snapshots of the state of the graph at different moments of its life. This object is often poorly used : most works focuss on the structure of each graph in the series. Doing so, one completely forget the relationships between the graphs of the series. We believe that these relationships encompass the essence of the structure of the dynamic and we place it at the very center of our approach. Thus, we put much effort on developping graph notions able to deal with a series of graphs instead of a dealing with a single graph. These notions must capture the temporal causality of the series and the non trivial relationships between its graphs. Our final goal is to provide a set of the notions and indicators to describe the dynamics of a network in a meaningful way, just like complex networks theory does for static complex networks.

Dynamic communities. The detection of dynamic communities is particularly appealing to describe dynamic networks. In order to extend the static case, one may apply existing community detection methods to successive snapshots of dynamic networks. This is however not totally satisfying for two main reasons: first, this would take a large amount of time (directly proportional to the data span); moreover, having a temporal succession of independent communities is not sufficient and we loose valuable information and dependencies. We also need to investigate the temporal links, study the time granularity and look for time periods that could be compressed within a single snapshot.

Tools for dynamic graph visualization. Designing generic and pure graph visualization tools is clearly out of the scope of the DANTE project. Efficient graph drawing tools or network analysis toolkit/software are now available (e.g., GUESS, TULIP, Sonivis, Network Workbench). However, the drawback of most softwares is that the dynamics is not taken into account. Since we will study the hierarchy of dynamics through the definition of communities we plan to extend graph drawing methods by using the communities' structures. We also plan to handle the time evolution in the network analysis toolkit. A tool like TULIP is well designed and could be improved by allowing operations (selection, grouping, sub graph computation...) to take place on the time dimension as well.

DIONYSOS Project-Team

3. Scientific Foundations

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 models' 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, ...), 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 measuring the QoE (see 3.2).

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 models' 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.

DISTRIBCOM Project-Team

3. Scientific Foundations

3.1. Overview of the needed paradigms

Management of telecommunications networks and services, and Web services, involves the following algorithmic tasks:

Observing, monitoring, and testing large distributed systems: Alarm or message correlation is one of the five basic tasks in network and service management. It consists in causally relating the various alarms collected throughout the considered infrastructure—be it a network or a service sitting on top of a transport infrastructure. Fault management requires in particular reconstructing the set of all state histories that can explain a given log of observations. Testing amounts to understanding and analyzing the responses of a network or service to a given set of stimuli; stimuli are generally selected according to given test purposes. All these are variants of the general problem of *observing* a network or service. Networks and services are large distributed systems, and we aim at observing them in a distributed way as well, namely: logs are collected in a distributed way and observation is performed by a distributed set of supervising peers.

Quality of Service (QoS) evaluation, negotiation, and monitoring: QoS issues are a well established topic for single domain networks or services, for various protocols — e.g., Diffserv for IP. Performance evaluation techniques are used that follow a "closed world" point of view: the modeling involves the overall traffic, and resource characteristics are assumed known. These approaches extend to some telecommunication services as well, e.g., when considering (G)MPLS over an IP network layer.

However, for higher level applications, including composite Web services (also called *orchestrations*), this approach to QoS is no longer valid. For instance, an orchestration using other Web services has no knowledge of how many users are calling the same Web services. In addition, it has no knowledge of the transport resources it is using. Therefore, the well developed "closed world" approach can no longer be used. *Contract-based* approaches are considered instead, in which a given orchestration offers promises to its users on the basis of promises it has from its subcontracting services. In this context, contract composition becomes a central issue. Monitoring is needed to check for possible breaching of the contract. Countermeasures would consist in reconfigurating the orchestration by replacing the failed subcontracted services by alternative ones.

The DistribCom team focuses on the algorithms supporting the above tasks. Therefore models providing an adequate framework are fundamental. We focus on models of discrete systems, not models of streams or fluid types of models. And we address the distributed and asynchronous nature of the underlying systems by using models involving only local, not global, states, and local, not global, time. These models are reviewed in section 3.2 . We use these mathematical models to support our algorithms and we use them also to study and develop formalisms of Web services orchestrations and workflow management in a more general setting.

3.2. Models of concurrency: nets, scenarios, event structures, and their variants

For Finite State Machines (FSM), a large body of theory has been developed to address problems such as: observation (the inference of hidden state trajectories from incomplete observations), control, diagnosis, and learning. These are difficult problems, even for simple models such as FSM's. One of the research tracks of DistribCom consists in extending such theories to distributed systems involving concurrency, i.e., systems in which both time and states are local, not global. For such systems, even very basic concepts such as "trajectories" or "executions" need to be deeply revisited. Computer scientists have for a long time recognized this topic of concurrent and distributed systems as a central one. In this section, we briefly introduce the reader to the models of scenarios, event structures, nets, languages of scenarios, graph grammars, and their variants.

3.2.1. Scenarios.

The simplest concept related to concurrency is that of a finite execution of a distributed machine. To this end, scenarios have been informally used by telecom engineers for a long time. In scenarios, so-called "instances" exchange asynchronous messages, thus creating events that are totally ordered on a given instance, and only partially ordered by causality on different instances (emission and reception of a message are causally related). The formalization of scenarios was introduced by the work done in the framework of ITU and OMG on High-level Message Sequence Charts and on UML Sequence Diagrams in the last ten years, see [52], [57]. This allowed in particular to formally define infinite scenarios, and to enhance them with variables, guards, etc [61], [59], [60]. Today, scenarios are routinely offered by UML and related software modeling tools.

3.2.2. Event structures.

Event structures were invented by Glynn Winskel and co-authors in 1980 [56], [62]. Executions are sets of events that are partially ordered by a *causality* relation. Event structures collect all the executions by superimposing shared prefixes. Events not belonging to a same execution are said in *conflict*. Events that are neither causally related nor in conflict are called *concurrent*. Concurrent processes model the "parallel progress" of components.

Categories of event structures have been defined, with associated morphisms, products, and co-products, see [63]. Products and co-products formalize the concepts of parallel composition and "union" of event structures, respectively. This provides the needed apparatus for composing and projecting (or abstracting) systems. Event structures have been mostly used to give the semantics of various formalisms or languages, such as Petri nets, CCS, CSP, etc [56], [62]. We in DistribCom make a nonstandard use of these, e.g., we use them as a structure to compute and express the solutions of observation or diagnosis problems, for concurrent systems.

3.2.3. Nets and languages of scenarios.

The next step is to have finite representations of systems having possibly infinite executions. In DistribCom, we use two such formalisms: *Petri nets* [58], [45] and *languages of scenarios* such as High-level Message Sequence Charts (HMSC) [52], [60]. Petri nets are well known, at least in their basic form, we do not introduce them here. We use so-called *safe* Petri Nets, in which markings are boolean (tokens can be either 0 or 1); and we use also variants, see below.

3.2.4. Extensions and variants.

Two extensions of the basic concepts of nets or scenario languages are useful for us. Nets or scenario languages enriched with variables, actions, and guards, are useful to model general concurrent and distributed dynamical systems in which a certain discrete abstraction of the control is represented by means of a net or a scenario language. Manipulating such *symbolic nets* requires using abstraction techniques. Time Petri nets and network of timed automata are particular cases of symbolic nets. Probabilistic Nets or event structures: Whereas a huge literature exists on stochastic Petri nets or stochastic process algebras (in computer science), randomizing *concurrent models,* i.e., with ω's being concurrent trajectories, not sequential ones, has been addressed only since the 21st century. We have contributed to this new area of research.

3.2.5. Handling dynamic changes in the systems.

The last and perhaps most important issue, for our applications, is the handling of dynamic changes in the systems model. This is motivated by the constant use of dynamic reconfigurations in management systems. Extensions of net models have been proposed to capture this, for example the *dynamic nets* of Vladimiro Sassone [44] and *net systems* [46]. For the moment, such models lack a suitable theory of unfoldings.

3.3. Modal logics for distributed systems

Modal logics are a family of logics that were developed originally to reason about different modalities occurring in natural language, such as for example the modality of knowledge (epistemic logic), the modalities of obligation and permission (deontic logic) and the modality of time (temporal logic). Temporal logics (CTL, LTL, μ-calculus...) are the most prominent (modal) logics used in computer science nowadays, especially in the field of verification.

3.3.1. Epistemic logic and distributed systems.

In the 1980's, epistemic logic was propounded by computer scientists such as Fagin, Halpern, Moses and Vardi to address problems in distributed systems, resulting in the TARK conference series (Theoretical Aspects of Rationality and Knowledge) and the books [48], [54]. This interest in epistemic logic was due to their observation that the notion of knowledge plays a central role in the informal reasoning used in the design of distributed protocols. This lead these authors to "hope that a theory of knowledge, communication and action will prove rich enough to provide general foundations for a unified theoretical treatment of distributed systems" [50]. The research pursued in DistribCom follows this line of thought, although we also strive to feed and confront our theoretical developments with actual problems stemming from diverse areas of application of distributed systems.

In [48], the behavior of a distributed system is represented by a set of *runs*, each run being a possible execution of the distributed system, determined by a given protocol. Processors are called *agents* and their partial observation of the system is represented at any point in the run by indistinguishability relations between local states of different runs (the local state of a processor represents the state of this processor at a moment of time). This model was used to show for example that the specific notion of common knowledge of epistemic logic is necessary to reach agreement and to coordinate actions [50]. Dynamic Epistemic Logic (DEL) is another logical framework that can be used to represent and reason about distributed systems (connections between these two logical frameworks were made in [64]). DEL deals with the representation of global states of synchronous distributed systems. The global state of the system at a moment in time is represented directly by means of an *epistemic model*. Events occurring in this distributed system are represented by means of *event models* and their effects on the local states of agents (processors) are represented by means of a *product update*.

The contributions in this sub-module are described in Section 6.4 .

3.3.2. Deontic logic and privacy in distributed systems.

We also use deontic logic in combination with epistemic logic for the formalization of privacy regulations. We intend to use this formalization to reason about privacy in the composition of web-services. The combination of these two modal logics can be used to express statements such as "it is *forbidden* for agent 1 to *know* that agent 2 sent message m" or "if agent 1 is an administrator of the system, then it is *permitted* for him to *know* information i". This provides a formal language very close to the natural language used in actual privacy regulations by law legislators. In the long run, we expect this formal language to be used at the level of interfaces of the web-service in order to:

1. check that the privacy policy declared by the web-service on its interface is indeed compliant (coherent) with respect to the privacy regulations expressed by law makers;
2. check that the web-service does enforce and apply the privacy policy it has declared on its interface.

The contributions in this sub-module are described in Section 6.8 .

3.4. Statistical Model Checking

Complex systems pose two particular challenges to formal verification: (i) the non-determinism caused by concurrency and unpredictable environmental conditions and (ii) the size of the state space. Our interest is probabilistic model checking, that can verify intricate details of a system's dynamical behavior and where non-determinism is handled by assigning probabilistic distributions to unknowns and quantifying results with a probability. Exact probabilistic model checking quantifies these probabilities to the limit of numerical precision by an exhaustive exploration of the state space, but is restricted by what can be conveniently stored in memory. Our focus is therefore statistical model checking (SMC), that avoids an explicit representation of the state space by building a statistical model of the executions of a system and giving results within confidence bounds. The key challenges of this approach are to reduce the length (simulation steps and cpu time) and number of simulation traces necessary to achieve a result with given confidence. Rare properties pose a particular problem in this respect, since they are not only difficult to observe but their probability is difficult to bound. A further goal is to make a tool where the choice of modeling language and logic are flexible.

FOCUS Project-Team

3. Scientific Foundations

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 Team

3. Scientific Foundations

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.

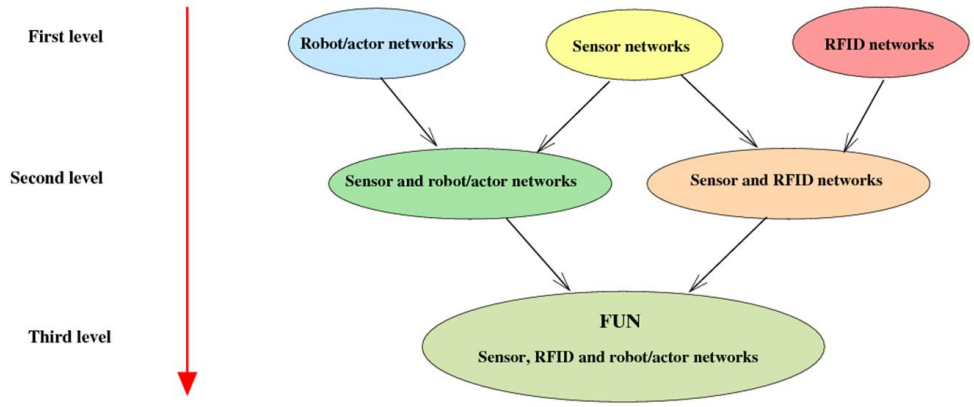


Figure 1. Panorama of FUN.

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 better measuring the influence of a realistic environment. 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 and Aspire [42]. Next steps will be to experiment not only on testbeds but also on real use cases. These latter will be given through different collaborations.

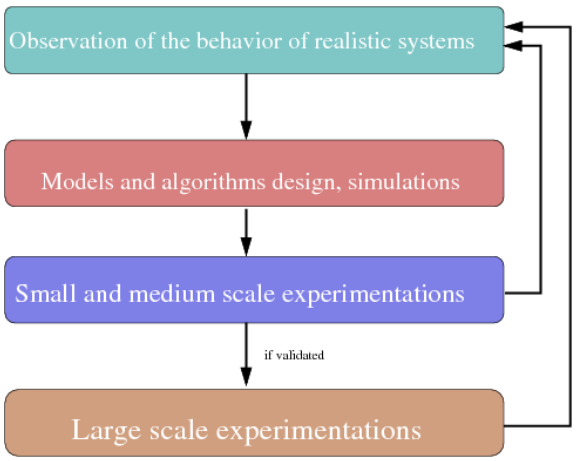


Figure 2. Methodology applied in the FUN research group.

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 brings 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 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 transparent merging of technologies. 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 [45]. 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 (section vide)

GRAND-LARGE Project-Team

3. Scientific Foundations

3.1. Large Scale Distributed Systems (LSDS)

What makes a fundamental difference between recent Global Computing systems (Seti@home), Grid (EGEE, TeraGrid) and former works on distributed systems is the large scale of these systems. This characteristic becomes also true for large scale parallel computers gathering tens of thousands of CPU cores. The notion of Large Scale is linked to a set of features that has to be taken into account in these systems. An example is the system dynamicity caused by node volatility: in Internet Computing Platforms (also called Desktop Grids), a non predictable number of nodes may leave the system at any time. Some recent results also report a very low MTTI (Mean Time To Interrupt) in top level supercomputers gathering 100,000+ CPU cores. Another example of characteristics is the complete lack of control of nodes connectivity. In Desktop Grid, we cannot assume that external administrator is able to intervene in the network setting of the nodes, especially their connection to Internet via NAT and Firewalls. This means that we have to deal with the in place infrastructure in terms of performance, heterogeneity, dynamicity and connectivity. These characteristics, associated with the requirement of scalability, establish a new research context in distributed systems. The Grand-Large project aims at investigating theoretically as well as experimentally the fundamental mechanisms of LSDS, especially for the high performance computing applications.

3.1.1. Computing on Large Scale Global Computing systems

Large scale parallel and distributed systems are mainly used in the context of Internet Computing. As a consequence, until Sept. 2007, Grand-Large has focused mainly on Desktop Grids. Desktop Grids are developed for computing (SETI@home, Folding@home, Decrypthon, etc.), file exchanges (Napster, Kazaa, eDonkey, Gnutella, etc.), networking experiments (PlanetLab, Porivo) and communications such as instant messaging and phone over IP (Jabber, Skype). In the High Performance Computing domain, LSDS have emerged while the community was considering clustering and hierarchical designs as good performance-cost tradeoffs. Nowadays, Internet Computing systems are still very popular (the BOINC platform is used to run over 40 Internet Computing projects and XtremWeb is used in production in three countries) and still raise important research issues.

Desktop Grid systems essentially extend the notion of computing beyond the frontier of administration domains. The very first paper discussing this type of systems [88] presented the Worm programs and several key ideas that are currently investigated in autonomous computing (self replication, migration, distributed coordination, etc.). LSDS inherit the principle of aggregating inexpensive, often already in place, resources, from past research in cycle stealing/resource sharing. Due to its high attractiveness, cycle stealing has been studied in many research projects like Condor [77] , Glunix [71] and Mosix [50], to cite a few. A first approach to cross administration domains was proposed by Web Computing projects such as Jet [81], Charlotte [51], Javeline [65], Bayanihan [86], SuperWeb [47], ParaWeb [58] and PopCorn [60]. These projects have emerged with Java, taking benefit of the virtual machine properties: high portability across heterogeneous hardware and OS, large diffusion of virtual machine in Web browsers and a strong security model associated with bytecode execution. Performance and functionality limitations are some of the fundamental motivations of the second generation of Global Computing systems like BOINC [49] and XtremWeb [67]. The second generation of Global Computing systems appeared in the form of generic middleware which allow scientists and programmers to design and set up their own distributed computing project. As a result, we have seen the emergence of large communities of volunteers and projects. Currently, Global Computing systems are among the largest distributed systems in the world. In the mean time, several studies succeeded to understand and enhance the performance of these systems, by characterizing the system resources in term of volatility and heterogeneity and by studying new scheduling heuristics to support new classes of applications: data-intensive, long running application with checkpoint, workflow, soft-real time etc... However, despite these

recent progresses, one can note that Global Computing systems are not yet part of high performance solution, commonly used by scientists. Recent researches to fulfill the requirements of Desktop Grids for high demanding users aim at redesigning Desktop Grid middleware by essentially turning a set of volatile nodes into a virtual cluster and allowing the deployment of regular HPC utilities (batch schedulers, parallel communication libraries, checkpoint services, etc...) on top of this virtual cluster. The new generation would permit a better integration in the environment of the scientists such as computational Grids, and consequently, would broaden the usage of Desktop Grid.

The high performance potential of LSDS platforms has also raised a significant interest in the industry. Performance demanding users are also interested by these platforms, considering their cost-performance ratio which is even lower than the one of clusters. Thus, several Desktop Grid platforms are daily used in production in large companies in the domains of pharmacology, petroleum, aerospace, etc.

Desktop Grids share with Grid a common objective: to extend the size and accessibility of a computing infrastructure beyond the limit of a single administration domain. In [68], the authors present the similarities and differences between Grid and Global Computing systems. Two important distinguishing parameters are the user community (professional or not) and the resource ownership (who own the resources and who is using them). From the system architecture perspective, we consider two main differences: the system scale and the lack of control of the participating resources. These two aspects have many consequences, at least on the architecture of system components, the deployment methods, programming models, security (trust) and more generally on the theoretical properties achievable by the system.

Beside Desktop Grids and Grids, large scale parallel computers with tens of thousands (and even hundreds of thousands) of CPU cores are emerging with scalability issues similar to the one of Internet Computing systems: fault tolerance at large scale, large scale data movements, tools and languages. Grand-Large is gradually considering the application of selected research results, in the domain of large scale parallel computers, in particular for the fault tolerance and language topics.

3.1.2. Building a Large Scale Distributed System

This set of studies considers the XtremWeb project as the basis for research, development and experimentation. This LSDS middleware is already operational. This set gathers 4 studies aiming at improving the mechanisms and enlarging the functionalities of LSDS dedicated to computing. The first study considers the architecture of the resource discovery engine which, in principle, is close to an indexing system. The second study concerns the storage and movements of data between the participants of a LSDS. In the third study, we address the issue of scheduling in LSDS in the context of multiple users and applications. Finally the last study seeks to improve the performance and reduce the resource cost of the MPICH-V fault tolerant MPI for desktop grids.

3.1.2.1. The resource discovery engine

A multi-users/multi-applications LSDS for computing would be in principle very close to a P2P file sharing system such as Napster [87], Gnutella [87] and Kazaa [76], except that the shared resource is the CPUs instead of files. The scale and lack of control are common features of the two kinds of systems. Thus, it is likely that solutions sharing fundamental mechanisms will be adopted, such as lower level communication protocols, resource publishing, resource discovery and distributed coordination. As an example, recent P2P projects have proposed distributed indexing systems like CAN [84], CHORD [89], PASTRY [85] and TAPESTRY [94] that could be used for resource discovery in a LSDS dedicated to computing.

The resource discovery engine is composed of a publishing system and a discovery engine, which allow a client of the system to discover the participating nodes offering some desired services. Currently, there is as much resource discovery architectures as LSDS and P2P systems. The architecture of a resource discovery engine is derived from some expected features such as speed of research, speed of reconfiguration, volatility tolerance, anonymity, limited use of the network, matching between the topologies of the underlying network and the virtual overlay network.

This study focuses on the first objective: to build a highly reliable and stable overlay network supporting the higher level services. The overlay network must be robust enough to survive unexpected behaviors (like malicious behaviors) or failures of the underlying network. Unfortunately it is well known that under specific assumptions, a system cannot solve even simples tasks with malicious participants. So, we focus the study on designing overlay algorithms for transient failures. A transient failure accepts any kind of behavior from the system, for a limited time. When failures stop, the system will eventually provide its normal service again.

A traditional way to cope with transient failures are self-stabilizing systems [66]. Existing self-stabilizing algorithms use an underlying network that is not compatible with LSDS. They assume that processors know their list of neighbors, which does not fit the P2P requirements. Our work proposes a new model for designing self-stabilizing algorithms without making this assumption, then we design, prove and evaluate overlay networks self-stabilizing algorithms in this model.

3.1.2.2. Fault Tolerant MPI

MPICH-V is a research effort with theoretical studies, experimental evaluations and pragmatic implementations aiming to provide a MPI implementation based on MPICH [79], featuring multiple fault tolerant protocols.

There is a long history of research in fault tolerance for distributed systems. We can distinguish the automatic/transparent approach from the manual/user controlled approach. The first approach relies either on coordinated checkpointing (global snapshot) or uncoordinated checkpointing associated with message logging. A well known algorithm for the first approach has been proposed by Chandy and Lamport [62]. This algorithm requires restarting all processes even if only one process crashes. So it is believed not to scale well. Several strategies have been proposed for message logging: optimistic [91], pessimistic [48], causal [92]. Several optimizations have been studied for the three strategies. The general context of our study is high performance computing on large platforms. One of the most used programming environments for such platforms is MPI.

Within the MPICH-V project, we have developed and published several original fault tolerant protocols for MPI: MPICH-V1 [55], MPICH-V2 [56], MPICH-Vcausal, MPICH-Vcl [57], MPICH-Pcl. The two first protocols rely on uncoordinated checkpointing associated with either remote pessimistic message logging or sender based pessimistic message logging. We have demonstrated that MPICH-V2 outperforms MPICH-V1. MPICH-Vcl implements a coordinated checkpoint strategy (Chandy-Lamport) removing the need of message logging. MPICH-V2 and Vcl are concurrent protocols for large clusters. We have compared them considering a new parameter for evaluating the merits of fault tolerant protocols: the impact of the fault frequency on the performance. We have demonstrated that the stress of the checkpoint server is the fundamental source of performance differences between the two techniques. MPICH-Vcausal implements a causal message logging protocols, removing the need for waiting acknowledgement in contrary to MPICH-V2. MPICH-Pcl is a blocking implementation of the Vcl protocol. Under the considered experimental conditions, message logging becomes more relevant than coordinated checkpoint when the fault frequency reaches 1 fault every 4 hours, for a cluster of 100 nodes sharing a single checkpoint server, considering a data set of 1 GB on each node and a 100 Mb/s network.

Multiple important events arose from this research topic. A new open source implementation of the MPI-2 standard was born during the evolution of the MPICH-V project, namely OpenMPI. OpenMPI is the result of the alliance of many MPI projects in the USA, and we are working to port our fault tolerance algorithms both into OpenMPI and MPICH.

Grids becoming more popular and accessible than ever, parallel applications developers now consider them as possible targets for computing demanding applications. MPI being the de-facto standard for the programming of parallel applications, many projects of MPI for the Grid appeared these last years. We contribute to this new way of using MPI through a European Project in which we intend to grid-enable OpenMPI and provide new fault-tolerance approaches fitted for the grid.

When introducing Fault-Tolerance in MPI libraries, one of the most neglected component is the runtime environment. Indeed, the traditional approach consists in restarting the whole application and runtime environment

in case of failure. A more efficient approach could be to implement a fault-tolerant runtime environment, capable of coping with failures at its level, thus avoiding the restart of this part of the application. The benefits would be a quicker restart time, and a better control of the application. However, in order to build a fault-tolerant runtime environment for MPI, new topologies, more connected, and more stable, must be integrated in the runtime environment.

For traditional parallel machines of large scale (like large scale clusters), we also continue our investigation of the various fault tolerance protocols, by designing, implementing and evaluating new protocols in the MPICH-V project.

3.2. Volatility and Reliability Processing

In a global computing application, users voluntarily lend the machines, during the period they don't use them. When they want to reuse the machines, it is essential to give them back immediately. We assume that there is no time for saving the state of the computation (for example because the user is shooting down is machine). Because the computer may not be available again, it is necessary to organize checkpoints. When the owner takes control of his machine, one must be able to continue the computation on another computer from a checkpoint as near as possible from the interrupted state.

The problems raised by this way of managing computations are numerous and difficult. They can be put into two categories: synchronization and repartition problems.

- Synchronization problems (example). Assume that the machine that is supposed to continue the computation is fixed and has a recent checkpoint. It would be easy to consider that this local checkpoint is a component of a global checkpoint and to simply rerun the computation. But on one hand the scalability and on the other hand the frequency of disconnections make the use of a global checkpoint totally unrealistic. Then the checkpoints have to be local and the problem of synchronizing the recovery machine with the application is raised.
- Repartition problems (example). As it is also unrealistic to wait for the computer to be available again before rerunning the interrupted application, one has to design a virtual machine organization, where a single virtual machine is implemented as several real ones. With too few real machines for a virtual one, one can produce starvation; with too many, the efficiency is not optimal. The good solution is certainly in a dynamic organization.

These types of problems are not new ([69]). They have been studied deeply and many algorithmic solutions and implementations are available. What is new here and makes these old solutions not usable is scalability. Any solution involving centralization is impossible to use in practice. Previous works validated on former networks can not be reused.

3.2.1. Reliability Processing

We voluntarily presented in a separate section the volatility problem because of its specificity both with respect to type of failures and to frequency of failures. But in a general manner, as any distributed system, a global computing system has to resist to a large set of failures, from crash failures to Byzantine failures, that are related to incorrect software or even malicious actions (unfortunately, this hypothesis has to be considered as shown by DECRYPTHON project or the use of erroneous clients in SETI@HOME project), with in between, transient failures such as loss of message duplication. On the other hand, failures related accidental or malicious memory corruptions have to be considered because they are directly related to the very nature of the Internet. Traditionally, two approaches (masking and non-masking) have been used to deal with reliability problems. A masking solution hides the failures to the user, while a non-masking one may let the user notice that failures occur. Here again, there exists a large literature on the subject (cf. [78], [90], [66] for surveys). Masking techniques, generally based on consensus, are not scalable because they systematically use generalized broadcasting. The self-stabilizing approach (a non-masking solution) is well adapted (specifically its time adaptive version, cf. [75], [74], [52], [53], [70]) for three main reasons:

1. Low overhead when stabilized. Once the system is stabilized, the overhead for maintaining correction is low because it only involves communications between neighbours.

2. Good adaptivity to the reliability level. Except when considering a system that is continuously under attacks, self-stabilization provides very satisfying solutions. The fact that during the stabilization phase, the correctness of the system is not necessarily satisfied is not a problem for many kinds of applications.
3. Lack of global administration of the system. A peer to peer system does not admit a centralized administrator that would be recognized by all components. A human intervention is thus not feasible and the system has to recover by itself from the failures of one or several components, that is precisely the feature of self-stabilizing systems.

We propose:

1. To study the reliability problems arising from a global computing system, and to design self-stabilizing solutions, with a special care for the overhead.
2. For problem that can be solved despite continuously unreliable environment (such as information retrieval in a network), to propose solutions that minimize the overhead in space and time resulting from the failures when they involve few components of the system.
3. For most critical modules, to study the possibility to use consensus based methods.
4. To build an adequate model for dealing with the trade-off between reliability and cost.

3.3. Parallel Programming on Peer-to-Peer Platforms (P5)

Several scientific applications, traditionally computed on classical parallel supercomputers, may now be adapted for geographically distributed heterogeneous resources. Large scale P2P systems are alternative computing facilities to solve grand challenge applications.

Peer-to-Peer computing paradigm for large scale scientific and engineering applications is emerging as a new potential solution for end-user scientists and engineers. We have to experiment and to evaluate such programming to be able to propose the larger possible virtualization of the underlying complexity for the end-user.

3.3.1. Large Scale Computational Sciences and Engineering

Parallel and distributed scientific application developments and resource managements in these environments are a new and complex undertaking. In scientific computation, the validity of calculations, the numerical stability, the choices of methods and software are depending of properties of each peer and its software and hardware environments; which are known only at run time and are non-deterministic. The research to obtain acceptable frameworks, methodologies, languages and tools to allow end-users to solve accurately their applications in this context is capital for the future of this programming paradigm.

GRID scientific and engineering computing exists already since more than a decade. Since the last few years, the scale of the problem sizes and the global complexity of the applications increase rapidly. The scientific simulation approach is now general in many scientific domains, in addition to theoretical and experimental aspects, often link to more classic methods. Several applications would be computed on world-spread networks of heterogeneous computers using some web-based Application Server Provider (ASP) dedicated to targeted scientific domains. New very strategic domains, such as Nanotechnologies, Climatology or Life Sciences, are in the forefront of these applications. The development in this very important domain and the leadership in many scientific domains will depend in a close future to the ability to experiment very large scale simulation on adequate systems [73]. The P2P scientific programming is a potential solution, which is based on existing computers and networks. The present scientific applications on such systems are only concerning problems which are mainly data independents: i.e. each peer does not communicate with the others.

P2P programming has to develop parallel programming paradigms which allow more complex dependencies between computing resources. This challenge is an important goal to be able to solve large scientific applications. The results would also be extrapolated toward future petascale heterogeneous hierarchically designed supercomputers.

3.3.2. Experimentations and Evaluations

We have followed two tracks. First, we did experiments on large P2P platforms in order to obtain a realistic evaluation of the performance we can expect. Second, we have set some hypothesis on peers, networks, and scheduling in order to have theoretical evaluations of the potential performance. Then, we have chosen a classical linear algebra method well-adapted to large granularity parallelism and asynchronous scheduling: the block Gauss-Jordan method to invert dense very large matrices. We have also chosen the calculation of one matrix polynomial, which generates computation schemes similar to many linear algebra iterative methods, well-adapted for very large sparse matrices. Thus, we were able to theoretically evaluate the potential throughput with respect to several parameters such as the matrix size and the multicast network speed.

Since the beginning of the evaluations, we experimented with those parallel methods on a few dozen peer XtremWeb P2P Platforms. We continue these experiments on larger platforms in order to compare these results to the theoretical ones. Then, we would be able to extrapolate and obtain potential performance for some scientific applications.

Recently, we also experimented several Krylov based method, such as the Lanczos and GMRES methods on several grids, such as a French-Japanese grid using hundred of PC in France and 4 clusters at the University of Tsukuba. We also experimented on GRID5000 the same methods. We currently use several middleware such as Xtremweb, OmniRPC and Condor. We also begin some experimentations on the Tsubame supercomputer in collaboration with the TITech (Tokyo Institute of Technologies) in order to compare our grid approaches and the High performance one on an hybrid supercomputer.

Experimentations and evaluation for several linear algebra methods for large matrices on P2P systems will always be developed all along the Grand Large project, to be able to confront the different results to the reality of the existing platforms.

As a challenge, we would like, in several months, to efficiently invert a dense matrix of size one million using a several thousand peer platform. We are already inverting very large dense matrices on Grid5000 but more efficient scheduler and a larger number of processors are required to this challenge.

Beyond the experimentations and the evaluations, we propose the basis of a methodology to efficiently program such platforms, which allow us to define languages, tools and interface for the end-user.

3.3.3. Languages, Tools and Interface

The underlying complexity of the Large Scale P2P programming has to be mainly virtualized for the end-user. We have to propose an interface between the end-user and the middleware which may extract the end-user expertise or propose an on-the-shelf general solution. Targeted applications concern very large scientific problems which have to be developed using component technologies and up-to-dated software technologies.

We introduced the YML framework and language which allows to describe dependencies between components. We introduced different classes of components, depending of the level of abstraction, which are associated with divers parts of the framework. A component catalogue is managed by an administrator and/or the end-users. Another catalogue is managed with respect to the experimental platform and the middleware criteria. A front-end part is completely independent of any middleware or testbed, and a back-end part is developed for each targeted middleware/platform couple. A YML scheduler is adapted for each of the targeted systems.

The YML framework and language propose a solution to develop scientific applications to P2P and GRID platform. An end-user can directly develop programs using this framework. Nevertheless, many end-users would prefer avoid programming at the component and dependency graph level. Then, an interface has to be proposed soon, using the YML framework. This interface may be dedicated to a special scientific domain to be able to focus on the end-user vocabulary and P2P programming knowledge. We plan to develop such version based on the YML framework and language. The first targeted scientific domain will be very large linear algebra for dense or sparse matrices.

3.4. Methodology for Large Scale Distributed Systems

Research in the context of LSDS involves understanding large scale phenomena from the theoretical point of view up to the experimental one under real life conditions.

One key aspects of the impact of large scale on LSDS is the emergence of phenomena which are not coordinated, intended or expected. These phenomena are the results of the combination of static and dynamic features of each component of LSDS: nodes (hardware, OS, workload, volatility), network (topology, congestion, fault), applications (algorithm, parameters, errors), users (behavior, number, friendly/aggressive).

Validating current and next generation of distributed systems targeting large-scale infrastructures is a complex task. Several methodologies are possible. However, experimental evaluations on real testbeds are unavoidable in the life-cycle of a distributed middleware prototype. In particular, performing such real experiments in a rigorous way requires to benchmark developed prototypes at larger and larger scales. Fulfilling this requirement is mandatory in order to fully observe and understand the behaviors of distributed systems. Such evaluations are indeed mandatory to validate (or not!) proposed models of these distributed systems, as well as to elaborate new models. Therefore, to enable an experimentally-driven approach for the design of next generation of large scale distributed systems, developing appropriate evaluation tools is an open challenge.

Fundamental aspects of LSDS as well as the development of middleware platforms are already existing in Grand-Large. Grand-Large aims at gathering several complementary techniques to study the impact of large scale in LSDS: observation tools, simulation, emulation and experimentation on real platforms.

3.4.1. Observation tools

Observation tools are mandatory to understand and extract the main influencing characteristics of a distributed system, especially at large scale. Observation tools produce data helping the design of many key mechanisms in a distributed system: fault tolerance, scheduling, etc. We pursue the objective of developing and deploying a large scale observation tool (XtremLab) capturing the behavior of thousands of nodes participating to popular Desktop Grid projects. The collected data will be stored, analyzed and used as reference in a simulator (SIMBOINC).

3.4.2. Tool for scalability evaluations

Several Grid and P2P systems simulators have been developed by other teams: SimGrid [61], GridSim [59], Briks [46]. All these simulators considers relatively small scale Grids. They have not been designed to scale and simulate 10 K to 100 K nodes. Other simulators have been designed for large multi-agents systems such as Swarm [80] but many of them considers synchronous systems where the system evolution is guided by phases. In the P2P field, ad hoc many simulators have been developed, mainly for routing in DHT. Emulation is another tool for experimenting systems and networks with a higher degree of realism. Compared to simulation, emulation can be used to study systems or networks 1 or 2 orders of magnitude smaller in terms of number of components. However, emulation runs the actual OS/middleware/applications on actual platform. Compared to real testbed, emulation considers conducting the experiments on a fully controlled platform where all static and dynamic parameters can be controlled and managed precisely. Another advantage of emulation over real testbed is the capacity to reproduce experimental conditions. Several implementations/configurations of the system components can be compared fairly by evaluating them under the similar static and dynamic conditions. Grand-Large is leading one of the largest Emulator project in Europe called Grid explorer (French funding). This project has built and used a 1K CPUs cluster as hardware platform and gathers 24 experiments of 80 researchers belonging to 13 different laboratories. Experiments concerned developing the emulator itself and use of the emulator to explore LSDS issues. In term of emulation tool, the main outcome of Grid explorer is the V-DS system, using virtualization techniques to fold a virtual distributed system 50 times larger than the actual execution platform. V-DS aims at discovering, understanding and managing implicit uncoordinated large scale phenomena. Grid Explorer is still in use within the Grid'5000 platform and serves the community of 400 users 7 days a week and 24h a day.

3.4.3. Real life testbeds: extreme realism

The study of actual performance and connectivity mechanisms of Desktop Grids needs some particular testbed where actual middleware and applications can be run under real scale and real life conditions. Grand-Large is

developing DSL-Lab, an experimental platform distributed on 50 sites (actual home of the participants) and using the actual DSL network as the connection between the nodes. Running experiments over DSL-Lab put the piece of software to study under extremely realistic conditions in terms of connectivity (NAT, Firewalls), performance (node and network), performance symmetry (DSL Network is not symmetric), etc.

To investigate real distributed system at large scale (Grids, Desktop Grids, P2P systems), under real life conditions, only a real platform (featuring several thousands of nodes), running the actual distributed system can provide enough details to clearly understand the performance and technical limits of a piece of software. Grand-Large members are strongly involved (as Project Director) in the French Grid5000 project which intents to deploy an experimental Grid testbed for computer scientists. This testbed features about 4000 CPUs gathering the resources of about 9 clusters geographically distributed over France. The clusters will be connected by a high speed network (Renater 10G). Grand-Large is the leading team in Grid5000, chairing the steering committee. As the Principal Investigator of the project, Grand-Large has taken some strong design decisions that nowadays give a real added value of Grid5000 compared to all other existing Grids: reconfiguration and isolation. From these two features, Grid5000 provides the capability to reproduce experimental conditions and thus experimental results, which is the cornerstone of any scientific instrument.

3.5. High Performance Scientific Computing

This research is in the area of high performance scientific computing, and in particular in parallel matrix algorithms. This is a subject of crucial importance for numerical simulations as well as other scientific and industrial applications, in which linear algebra problems arise frequently. The modern numerical simulations coupled with ever growing and more powerful computational platforms have been a major driving force behind a progress in numerous areas as different as fundamental science, technical/technological applications, life sciences.

The main focus of this research is on the design of efficient, portable linear algebra algorithms, such that solving a large set of linear equations or a least squares problem. The characteristics of the matrices commonly encountered in this situations can vary significantly, as are the computational platforms used for the calculations. Nonetheless two common trends are easily discernible. First, the problems to solve are larger and larger, since the numerical simulations are using higher resolution. Second, the architecture of today's supercomputers is getting very complex, and so the developed algorithms need to be adapted to these new achitectures.

3.5.1. Communication avoiding algorithms for numerical linear algebra

Since 2007, we work on a novel approach to dense and sparse linear algebra algorithms, which aims at minimizing the communication, in terms of both its volume and a number of transferred messages. This research is motivated by technological trends showing an increasing communication cost. Its main goal is to reformulate and redesign linear algebra algorithms so that they are optimal in an amount of the communication they perform, while retaining the numerical stability. The work here involves both theoretical investigation and practical coding on diverse computational platforms. We refer to the new algorithms as *communication avoiding algorithms* [18] [8]. In our team we focus on communication avoiding algorithms for dense direct methods as well as sparse iterative methods.

The theoretical investigation focuses on identifying lower bounds on communication for different operations in linear algebra, where communication refers to data movement between processors in the parallel case, and to data movement between different levels of memory hierarchy in the sequential case. The lower bounds are used to study the existing algorithms, understand their communication bottlenecks, and design new algorithms that attain them.

This research focuses on the design of linear algebra algorithms that minimize the cost of communication. Communication costs include both latency and bandwidth, whether between processors on a parallel computer or between memory hierarchy levels on a sequential machine. The stability of the new algorithms represents an important part of this work.

3.5.2. Preconditioning techniques

Solving a sparse linear system of equations is the most time consuming operation at the heart of many scientific applications, and therefore it has received a lot of attention over the years. While direct methods are robust, they are often prohibitive because of their time and memory requirements. Iterative methods are widely used because of their limited memory requirements, but they need an efficient preconditioner to accelerate their convergence. In this direction of research we focus on preconditioning techniques for solving large sparse systems.

One of the main challenges that we address is the 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 filtering LU decompositions. A judicious choice for the directions to be preserved through filtering allows us to alleviate the effect of low frequency modes on the convergence. While preconditioners and their scalability are studied by many other groups, our approach of direction preserving and filtering is studied in only very few other groups in the world (as Lawrence Livermore National Laboratory, Frankfurt University, Pennsylvania State University).

3.5.3. Fast linear algebra solvers based on randomization

Linear algebra calculations can be enhanced by statistical techniques in the case of a square linear system $Ax = b$ where A is a general or symmetric indefinite matrix [16]& [25]. Thanks to a random transformation of A, it is possible to avoid pivoting and then to reduce the amount of communication. Numerical experiments show that this randomization can be performed at a very affordable computational price while providing us with a satisfying accuracy when compared to partial pivoting. This random transformation called Partial Random Butterfly Transformation (PRBT) is optimized in terms of data storage and flops count. A PRBT solver for LU factorization (and for LDL^T factorization on multicore) has been developed. This solver takes advantage of the latest generation of hybrid multicore/GPU machines and gives better Gflop/s performance than existing factorization routines.

3.5.4. Sensitivity analysis of linear algebra problems

We derive closed formulas for the condition number of a linear function of the total least squares solution [2]. Given an over determined linear systems $Ax = b$, we show that this condition number can be computed using the singular values and the right singular vectors of $[A, b]$ and A. We also provide an upper bound that requires the computation of the largest and the smallest singular value of $[A, b]$ and the smallest singular value of A. In numerical experiments, we compare these values with condition estimates from the literature.

HIEPACS Project-Team

3. Scientific Foundations

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 its 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 millions of cores. 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 will be the basis of both academic and industrial collaborations described in Section 4.2 and Section 4.3 , 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 modelling 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 .

3.2. High-performance computing on next generation architectures

Participants: Rached Abdelkhalek, Emmanuel Agullo, Olivier Coulaud, Iain Duff, Pierre Fortin, Luc Giraud, Abdou Guermouche, Andra Hugo, Guillaume Latu, Stojce Nakov, 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 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 (i.e., Blue Gene-like platforms) 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 problem. This has to be done in collaboration with our colleagues from the scheduling community like for example O. Beaumont (Inria CEPAGE 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 done 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 `marcel` thread library 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 (like clearspeed cards, ...), 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, in the context of the PhD of Andra Hugo, we investigate new approaches for composing parallel applications within a runtime system for heterogeneous platforms. The main goal of this work is to build an improved runtime system which is able to deal with parallel tasks (which may use different parallelization schemes or even different parallelization supports). More precisely, together with members from the Inria Runtime project-team, we proposed an extension of `StarPU`, a runtime system specifically designed for heterogeneous architectures, that allows multiple parallel codes to run concurrently with minimal interference. Such parallel codes run within *scheduling contexts* that provide confined execution environments which can be used to partition computing resources. Scheduling contexts can be dynamically resized to optimize the allocation of computing resources among concurrently running libraries. We introduced a *hypervisor* that automatically expands or shrinks contexts using feedback from the runtime system (e.g. resource utilization). We demonstrated the relevance of our approach using benchmarks invoking multiple high performance linear algebra kernels simultaneously on top of heterogeneous multicore machines. We showed that our mechanism can dramatically improve the overall application run time (-34%), most notably by reducing the average cache miss ratio (-50%).

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 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. The objectives will be to develop dense linear algebra algorithms and libraries for multicore architectures in the context the PLASMA project (http://icl.cs.utk.edu/plasma/) and for GPU and hybrid multicore/GPU architectures in the context of the MAGMA project (http://icl.cs.utk.edu/magma/). The framework that hosts all these research activities is the associated team MORSE (http://www.inria.fr/en/teams/morse).

A more prospective objective is to study the fault tolerance 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 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 done 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 GRAAL and GRAND-LARGE. 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.

3.3. High performance solvers for large linear algebra problems

Participants: Emmanuel Agullo, Olivier Coulaud, Iain Duff, Luc Giraud, Abdou Guermouche, Andra Hugo, Yan-Fei Jing, Jean Roman, Pablo Salas Medina, Stojce Nakov, Xavier Vasseur, Mawussi Zounon.

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 such 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. Although we will not contribute directly to this activity, we will use parallel sparse direct solvers as 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.

3.3.1. 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 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 attempt to apply to general unstructured linear systems domain decomposition ideas. More precisely, we will consider numerical techniques based on a non-overlapping decomposition of the graph associated with the sparse matrices. The vertex separator, built by a graph partitioner, will define the interface variables that will be solved iteratively using a Schur complement techniques, while the variables associated with the internal sub-graphs will be handled by a sparse direct solver. Although the Schur complement system is usually more tractable than the original problem by an iterative technique, preconditioning treatment is still required. For that purpose, the algebraic additive Schwarz technique initially developed for the solution of linear systems arising from the discretization of elliptic and parabolic PDE's will be extended. Linear systems where the associated matrices are symmetric in pattern will be first studied but extension to unsymmetric matrices will be latter considered. The main focus will be on difficult problems (including non-symmetric and indefinite ones) where it is harder to prevent growth in the number of iterations with the number of subdomains when considering massively parallel platforms. In that respect, we will consider algorithms that exploit several sources and grains of parallelism to achieve high computational throughput. This activity may involve collaborations with developers of sparse direct solvers as well as with developers of run-time systems and will lead to the development to the library `MaPHyS` (see Section 5.2). Some specific aspects, such as mixed MPI-thread implementation for the computer science aspects and techniques for indefinite system for the numerical aspects will be investigated in the framework of a France Berkeley Fund project granted that started last year.

3.3.2. 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 sides 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. 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 built augmented Krylov subspaces. In that context, Yan-Fei Jing, who joint `HiePACS`as post-doc, is investigating how reliable block Arnoldi procedure can be combined with deflated restarted block GMRES technique.

Flexible Krylov subspace methods with recycling techniques. In many situations, it has been observed that significant convergence improvements can be achieved in preconditioned Krylov subspace methods by enriching them with some spectral information. On the other hand effective preconditioning strategies are often designed where the preconditioner varies from one step to the next (e.g. in domain decomposition

methods, when approximate solvers are considered for the interior problems, or more generally for block preconditioning technique where approximate block solution are used) so that a flexible Krylov solver is required. In that context, we intend to investigate how numerical techniques implementing subspace recycling and/or incremental preconditioning can be extended and adapted to cope with this situation of flexible preconditioning; that is, how can we numerically benefit from the preconditioning implementation flexibility.

Krylov solver for complex symmetric non-Hermitian matrices. In material physics when the absorption spectrum of a molecule due to an exterior field is computed, we have to solve for each frequency a dense linear system where the matrix depends on the frequency. The sequence of matrices are complex symmetric non-Hermitian. While a direct approach can be used for small molecules, a Krylov subspace solver must be considered for larger molecules. Typically, Lanczos-type methods are used to solve these systems but the convergence is often slow. Based on our earlier experience on preconditioning techniques for dense complex symmetric non-Hermitian linear system in electromagnetism, we are interested in designing new preconditioners for this class of material physics applications. A first track will consist in building preconditioners on sparsified approximation of the matrix as well as computing incremental updates, eg. Sherman-Morrison type, of the preconditioner when the frequency varies. This action will be developed in the framework of the research activity described in Section 4.2 .

Approximate factoring of the inverse. When the matrix of a given sparse linear system of equations is known to be nonsingular, the computation of approximate factors for the inverse constitutes an algebraic approach to preconditioning. The main aim is to combine standard preconditioning ideas with sparse approximate inverse approximation to have implicitly dense approximate inverse approximations. Theory has been developed and encouraging numerical experiments have been obtained on a set of sparse matrices of small to medium size. We plan to propose a parallel implementation of the construction of the preconditioner and to investigate its efficiency on real-life problems. Extension of this technique to build a sparse approximation of the Schur complement for algebraic domain decomposition has also been investigated and could be integrated in the `MaPHySpackage` in the future.

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.

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 three 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 framework of an FP7 Marie project (MyPlanet), we intend to study parallel robust nonlinear quadratic eigensolvers. It is a crucial question in numerous technologies like the stability and vibration analysis in classical structural mechanics. The first research action consists in enhancing the robustness of the linear eigensolver and to consider shift invert technique to tackle difficult problems out of reach with the current technique. One of the main constraint in that framework is to design matrix-free technique to limit the memory consumption of the complete solver. For the nonlinear part different approaches ranging from simple nonlinear stationary iterations to Newton's type approaches will be considered.

- 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 (see Section 4.2).

3.4. High performance Fast Multipole Method for N-body problems

Participants: Emmanuel Agullo, Bérenger Bramas, Arnaud Etcheverry, Olivier Coulaud, Pierre Fortin, Luc Giraud, Matthias Messner, Jean Roman.

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 error tolerance. 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. In the former `ScAlApplix` project, 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 acheive high performance. Funded by the ANR-OPTIDIS, Arnaud Etcheverry started a PhD in October to study parallel scalable FMM techniques for the dislocation calculations.

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 algorithmics for code coupling in complex simulations

Participants: Olivier Coulaud, Aurélien Esnard, Jean Roman, Jérôme Soumagne, Clément Vuchener.

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 are 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.

The experience that we have acquired in the `ScAlApplix` project through the activities in crack propagation simulations with LibMultiScale and in M-by-N computational steering (coupling simulation with parallel visualization tools) with `EPSN` shows us that if the model aspect was well studied, several problems in parallel or distributed algorithms are still open and not well studied. In the context of code coupling in `HiePACS` we want to contribute more precisely to the following points.

3.5.1. Efficient schemes for multiscale simulations

As mentioned previously, many important physical phenomena, such as material deformation and failure (see Section 4.2), are inherently multiscale processes that cannot always be modeled via continuum model. Fully microspcopic 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 (ANR), 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 NOSSI ANR and the atomic to dislocation coupling in the OPTIDIS ANR (proposal for the 2010 COSINUS call of the French ANR).

3.5.2. Load-balancing of complex coupled simulations based on the hypergraph model

One most important issue is undoubtedly the problem of load-balancing of the whole coupled simulation. Indeed, the naive balancing of each code on its own can lead to important imbalance in the coupling area. Another connected problem we plan to investigate is the problem of resource allocation. This is particularly important for the global coupling efficiency, 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 codes to avoid that one of them wait for the others.

The performance of the coupled codes depends on how the data are well distributed on the processors. Generally, the data distributions of each code are built independently from each other to obtain the best load-balancing. But once the codes are coupled, the naive use of these decompositions can lead to important imbalance in the coupling area. Therefore, the modeling of the whole coupling is crucial to improve the performance and to ensure a good scalability. The goal is to find the best data distribution for the whole coupled codes and not only for each standalone code. One idea is to use an hypergraph model that will incorporate information about the coupling itself. Then, we expect the greater expressiveness of hypergraph will enable us to perform a coupling-aware partitioning in order to improve the load-balancing of the whole coupled simulation.

Another connected problem we plan to investigate is the problem of resource allocation. This is particularly important for the global coupling efficiency and scalability, 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 codes to avoid that one of them wait for the others. Typically, if we have a given number of processors and two coupled codes, how to split the processors among each code?

Moreover, the load-balancing of modern parallel adaptive simulations raises a crucial issue when the problem size varies during execution. In such cases, it could be convenient to dynamically adapt the number of resources used at runtime. However, most of previous works on repartitioning only consider a constant number of resources. We plan to design new repartitioning algorithm based on an hypergraph model that can handle a variable number of processors. Furthermore, this kind of algorithms could be used for the dynamic balancing of a coupled simulation, in the case where the whole number of resources is fixed but can change for each code.

3.5.3. Steering and interacting with complex coupled simulations

The computational steering is an effort to make the typical simulation work-flow (modelling, computing, analyzing) more efficient, by providing online visualization and interactive steering over the on-going computational processes. The online visualization appears very useful to monitor and to detect possible errors in long-running applications, and the interactive steering allows the researcher to alter simulation parameters on-the-fly and to immediately receive feedback on their effects. Thus, the scientist gains an additional insight in the simulation regarding to the cause-and-effect relationship.

In the `ScAlApplix` project, we have studied this problem in the case where both the simulation and the visualization can be parallel, what we call M-by-N computational steering, and we have developed a software environment called `EPSN` (see Section 5.3). More recently, we have proposed a model for the steering of complex coupled simulations and one important conclusion we have from these previous works is that the steering problem can be conveniently modeled as a coupling problem between one or more parallel simulation codes and one visualization code, that can be parallel as well. We propose in `HiePACS` to revisit the steering problem as a coupling problem and we expect to reuse the new redistribution algorithms developped in the context of code coupling for the purpose of M-by-N steering. We expect such an approach will enable to steer massively-parallel simulations. Another point we plan to study is the monitoring and interaction with resources, in order to perform user-directed checkpoint/restart or user-directed load-balancing at runtime.

In several applications, it is often very useful either to visualize the results of the ongoing simulation before writing it to disk, or to steer the simulation by modifying some parameters and visualize the impact of these modifications interactively. Nowadays, high performance computing simulations use many computing nodes, that perform I/O using the widely used HDF5 file format. One of the problems is now to use real-time visualization using high performance computing. In that respect we need to efficiently combine very large parallel simulation systems with parallel visualization systems. The originality of this approach is the use of the HDF5 file format to write in a distributed shared memory (DSM); so that the data can be read from the upper part of the visualization pipeline. This leads to define a relevant steering model based on a DSM. It implies finding a way to write/read data efficiently in this DSM, and steer the simulation. This work is developed in collaboration with the Swiss National Supercomputing Centre (CSCS).

As concerns the interaction aspect, we are interested in providing new mechanisms to interact with the simulation directly through the visualization. For instance in the ANR NOSSI, in order to speed up the computation we are interested in rotating a molecule in a cavity or in moving it from one cavity to another within the crystal latice. To perform safely such interactions a model of the interaction in our steering framework is necessary to keep the data coherency in the simulation. Another point we plan to study is the monitoring and interaction with ressources, in order to perform user-directed checkpoint/restart or user-directed load balancing at runtime.

HIPERCOM Project-Team

3. Scientific Foundations

3.1. Analytical information theory

Participants: Cédric Adjih, Pascale Minet, Paul Mühlethaler.

channel capacity, compression, predictors

Information theory Branch of mathematics dedicated to the quantification of the performance of a medium to carry information. Initiated by Shannon in 1948.

Abstract. Information theory and analytical methods play a central role in the networking technology. It identifies the key parameter that must be quantified in order to characterize the performance of a network.

The analytical information theory is part of the foundations of the Hipercom project. This is a tool box that has been collected and adapted from the areas of the analysis of algorithms and the information theory. It provides powerful tool for the analysis of telecommunication algorithms. The analysis of the behavior of such algorithms in their asymptotic range are fundamental in order to identify their critical parts. It helps to design and properly scale the protocols. Application of analytical information theory ranges from channel capacity computations, compression algorithm performance evaluation, predictor designs.

3.2. Methodology of telecommunication algorithm evaluation

Participants: Cédric Adjih, Ichrak Amdouni, Emmanuel Baccelli, Salman Malik, Yacine Mezali, Pascale Minet, Paul Mühlethaler, Saoucene Mahfoudh Ridene, Ridha Soua, Erwan Livolant, Ines Khoufi.

deterministic performance, probabilistic performance

Power laws probability distributions that decays has inverse power of the variable for large values of the variable. Power laws are frequent in economic and statistical analysis (see Pareto law). Simple models such as Poisson processes and finite state Markov processes don't generate distributions with power laws.

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 to identify and quantify 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 systemized 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.

3.3. Traffic and network architecture modeling

Participants: Cédric Adjih, Aline Carneiro Viana, Emmanuel Baccelli.

traffic source models, network topologies, mobility models, dynamic nodes
Abstract. Network models are important. We consider four model problems: topology, mobility, dynamics and traffic models.

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 a coverage area, or because some other nodes leaves the coverage area.

The traffic model is very most important. There are plenty 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 dependences that don't exist in Poisson traffic. There are still strong issues about the origin of this long range dependences 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.

3.4. Algorithm design, evaluation and implementation

Participants: Cédric Adjih, Aline Carneiro Viana, Emmanuel Baccelli, Saoucene Mahfoudh Ridene, Pascale Minet, Paul Mühlethaler, Ridha Soua, Erwan Livolant, Ines Khoufi.

Access protocols, routing, scheduling, QoS

Abstract. Algorithms are conceived with focal point on performance. The algorithms we specify in detail range between medium access control to admission control and quality of service management.

The conception of algorithms is an important focus of the project 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 imbedded simulation with existing network simulators. The algorithm specified by the project ranges from multiple access schemes, wireless ad hoc routing, mobile multicast management, Quality of service and admission controls. 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 be performing such to 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 are found are quantified with respect to the optimal solution.

INDES Project-Team

3. Scientific Foundations

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 [4]. 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 [5]. 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 [3]. 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.

KERDATA Project-Team

3. Scientific Foundations

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 [32] parallel programming paradigm, currently used on most cloud platforms, following the trend set up by Amazon [28]. At the core of the Map-Reduce frameworks stays 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 highlights a major, more general research topic. It has been clearly identified by international panels of experts like IESP [30] and EESI [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 infrastructure and methodology 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. We are also looking at other kinds of infrastructures (that we are considering as secondary), 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, University of Tsukuba) and Industry (Microsoft, IBM).

The highly experimental nature of our research validation methodology should be stressed. Our approach relies on building prototypes and on their large-scale experimental validation on real testbeds and experimental platforms. We strongly rely on the ALADDIN-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, hopefully soon, to the Blue Waters supercomputer). This provides us with excellent opportunities to validate our results on 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 potential impact.

3.2. Our research agenda

Three typical application scenarios are described in Section 4.1 :

- 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 efficiently address the aforementioned challenges 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 more in depth two usage scenarios, which led to two more specific approaches: 1) Pyramid (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 this approach in preliminary experiments [37], we have initiated several collaborative projects led by KerData in the area of cloud data management, e.g., the MapReduce ANR project, the A-Brain 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 those projects, we have already started to enjoy a visible scientific positioning at the international level.

The particularly active DataCloud@work 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 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 .

LOGNET Team (section vide)

MADYNES Project-Team

3. Scientific Foundations

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. Scientific Foundations

3.1. Scientific Foundations

The main mathematical tools and formalisms used in MAESTRO include:

- theory of stochastic processes: Markov process, renewal process, point process, Palm measure, large deviations, branching process, mean-field approximation;
- theory of dynamical discrete-event systems: queues, fluid approximation;
- theory of control and scheduling: dynamic programming, Markov decision process, game theory, deterministic and stochastic scheduling, pathwise comparison;
- theory of singular perturbations;
- random matrix theory.

MASCOTTE Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

The project develops tools and theory in the following domains: Discrete Mathematics (in particular Graph Theory), Algorithmics, Combinatorial Optimization and Simulation.

Typically, a telecommunication network (or an interconnection network) is modeled by a graph. A vertex may represent either a processor or a router or any of the following: a switch, a radio device, a site or a person. An edge (or arc) corresponds to a connection between the elements represented by the vertices (logical or physical connection). We can associate more information both to the vertices (for example what kind of switch is used, optical or not, number of ports, equipment cost) and to the edges (weights which might correspond to length, cost, bandwidth, capacity) or colors (modeling either wavelengths or frequencies or failures) etc. Depending on the application, various models can be defined and have to be specified. This modeling part is an important task. To solve the problems, we manage, when possible, to find polynomial algorithms. For example, a maximum set of disjoint paths between two given vertices is by Menger's theorem equal to the minimum cardinality of a cut. This problem can be solved in polynomial time using graph theoretic tools or flow theory or linear programming. On the contrary, determining whether in a directed graph there exists a pair of disjoint paths, one from s_1 to t_1 and the other from s_2 to t_2, is an NP-complete problem, and so are all the problems which aim at minimizing the cost of a network which can satisfy certain traffic requirements. In addition to deterministic hypotheses (for example if a connection fails it is considered as definitely down and not intermittently), the project started recently to consider probabilistic ones.

Graph coloring is an example of concept which appears in various contexts: WDM networks where colors represent wavelengths, radio networks where colors represent frequencies, fault tolerance where colors represent shared risk resource groups, and scheduling problems. Another tool concerns the development of new algorithmic aspects like parameterized algorithms.

MESCAL Project-Team

3. Scientific Foundations

3.1. Large System Modeling and Analysis

Participants: Bruno Gaujal, Derrick Kondo, Arnaud Legrand, Panayotis Mertikopoulos, Florence Perronnin, Brigitte Plateau, Olivier Richard, Corinne Touati, 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 infeasible 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 not only enable one to get repeatable results but also make it possible 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 a 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 tools 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. Equilibria can also be improved by advising policies to mobiles such that any user that does not follow these pieces of advice will necessarily penalize herself (*correlated equilibria*).

3.2. Management of Large Architectures

Participants: Derrick Kondo, Arnaud Legrand, Olivier Richard, Corinne Touati.

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 os 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 of old-fashioned conception, 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 requests 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 and 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. The reliability of the system is hence a crucial point. 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. The first one relies 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 much higher level nature.

In order to use large data, it is necessary (but not always sufficient, as seen later) to efficiently store and transfer them to a given site (a set of nodes) where it is going to be used. The first step toward this achievement is the construction of a file system that is an extension of NFS for the grid environment. The second step is an efficient transfer tool that provides throughput close to optimal (*i.e.* the capacity of the underlying hardware).

Our goal here is to design a distributed file system for clusters that enables one to store data over a set of nodes (instead of a single one). It was designed to permit the usage of a set of disks to optimize memory allocations. It is important for performance and simplicity that this new file system has little overhead for access and updates. From a user point of view, it is used just as a classical NFS. From the server point of view, however, the storage is distributed over several nodes (possibly including the users).

MOAIS Project-Team

3. Scientific Foundations

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 **Kaapi**.

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 Kaapi (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. Finally, to validate these novel parallel computation schemes, we developed a tool named **KRASH**. This tool is able to generate dynamic CPU load in a reproducible way on many-cores machines. Thus, by providing the same experimental conditions to several parallel applications, it enables users to evaluate the efficiency of resource uses for each approach.

By optimizing the work-stealing to our adaptive algorithm scheme, the non-blocking (wait-free) implementation of Kaapi has been designed and leads to the C library X-kaapi.

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.

MYRIADS Project-Team

3. Scientific Foundations

3.1. Introduction

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 [68] 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 companies 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 computer 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.

Clouds can be defined as platforms for on-demand resource provisioning over the Internet. These platforms rely on networked computers. Three flavours of cloud platforms have emerged corresponding to different kinds of service delivery:

- IaaS (Infrastructure as a Service) refers to clouds for on-demand provisioning of elastic and customizable execution platforms (from physical to virtualized hardware).
- PaaS (Platform as a Service) refers to clouds providing an integrated environment to develop, build, deploy, host and maintain scalable and adaptable applications.
- SaaS (Software as a Service) refers to clouds providing customers access to ready-to-use applications.

The cloud computing model [65], [62] introduces new challenges in the organization of the information infrastructure: security, identity management, adaptation to the environment (costs). The organization of large organization 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 will work on the design and implementation of system services to autonomously manage cloud and cloud federations resources and support collaboration between cloud users.

3.5. Unconventional/Nature-inspired Programming

Facing the complexity of the emerging ICT landscape in which highly heterogeneous digital services evolve and interact in numerous different ways in an autonomous fashion, there is a strong need for rethinking programming models. The question is "*what programming paradigm can efficiently and naturally express this great number of interactions arising concurrently on the platform?*.

It has been suggested [63] that observing nature could be of great interest to tackle the problem of modeling and programming complex computing platforms, and overcome the limits of traditional programming models. Innovating unconventional programming paradigms are requested to provide a high-level view of these interactions, then allowing to clearly separate what is a matter of expression from what is a question of implementation. Towards this, nature is of high inspiration, providing examples of self-organising, fully decentralized coordination of complex and large scale systems.

As an example, chemical computing [66] has been proposed more than twenty years ago for a natural way to program parallelism. Even after significant spread of this approach, it appears today that chemical computing exposes a lot of good properties (implicit autonomy, decentralization, and parallelism) to be leveraged for programming service infrastructures.

The Myriads team will investigate nature-inspired programming such as chemical computing for autonomous service computing.

OASIS Project-Team

3. Scientific Foundations

3.1. Programming with distributed objects and components

The paradigm of object-oriented programming, although not very recent, is clearly still not properly defined and implemented; for example notions like inheritance, sub-typing or overloading have as many definitions as there are different object languages. The introduction of concurrency and distribution into objects also increases the complexity. It appeared that standard Java constituents such as RMI (Remote Method Invocation) do not help building, in a transparent way, sequential, multi-threaded, or distributed applications. Indeed allowing, as RMI does, the execution of the same application to proceed on a shared-memory multiprocessors architecture as well as on a network of computing units (intranet, Internet), or on any hierarchical combination of both, is not sufficient for providing a convenient and reliable programming environment.

The question is thus: how to ease the construction (i.e. programming), deployment and evolution of distributed applications ?

One of the answers we suggest relies on the concept of active object, that acts as a single entity, abstraction of a thread, a set of objects and a location. Active objects communicate by asynchronous method calls thanks to the use of futures. ProActive is a Java library that implements this notion of active objects. ProActive can also be seen as a middleware supporting deployment, runtime support, and efficient communication for large scale distributed applications.

Another answer we provide relies on component-oriented programming. In particular, we have defined parallel and hierarchical distributed components starting from the Fractal component model developed by Inria and France-Telecom [58]. We have been involved in the design of the Grid Component Model (GCM) [4], which is one of the major results produced by the CoreGrid European Network of Excellence. The GCM has been standardized at ETSI ([64] for the last published standard), and most of our research on component models are related to it. On the practical side, ProActive/GCM is the implementation of the GCM above the ProActive programming library.

We have developed over time skills in both theoretical and applicative side fields, such as distribution, fault-tolerance, verification, etc., to provide a better programming and runtime environment for object oriented and component oriented applications.

3.2. Formal models for distributed objects

A few years ago, we designed the ASP calculus [7] for modelling distributed objects. It remains to this date one of our major scientific foundations. ASP is a calculus for distributed objects interacting using asynchronous method calls with generalized futures. Those futures naturally come with a transparent and automatic synchronisation called wait-by-necessity. In large-scale systems, our approach provides both a good structure and a strong decoupling between threads, and thus scalability. Our work on ASP provides very generic results on expressiveness and determinism, and the potential of this approach has been further demonstrated by its capacity to cope with advanced issues, such as mobility, group communications, and components [6].

ASP provides confluence and determinism properties for distributed objects. Such results should allow one to program parallel and distributed applications that behave in a deterministic manner, even if they are distributed over local or wide area networks.

The ASP calculus is a model for the ProActive library. An extension of ASP has been built to model distributed asynchronous components. A functional fragment of ASP has been modelled in the Isabelle theorem prover [8].

3.3. Verification, static analysis, and model-checking

Even with the help of high-level libraries, distributed systems are more difficult to program than classical applications. The complexity of interactions and synchronisations between remote parts of a system increases the difficulty of analysing their behaviours. Consequently, safety, security, or liveness properties are particularly difficult to ensure for these applications. Formal verification of software systems has been active for a long time, but its impact on the development methodology and tools has been slower than in the domain of hardware and circuits. This is true both at a theoretical and at a practical level; our contributions include:

- the definition of adequate models representing programs,
- the mastering of state complexity through abstraction techniques, new algorithmic approaches, or research on advanced parallel or distributed verification methods,
- the design of software tools that hide to the final user the complexity of the underlying theory.

We concentrate on the area of distributed component systems, where we get better descriptions of the structure of the system, making the analysis more tractable, but we also find out new interesting problems. For instance, we contributed to a better analysis of the interplay between the functional definition of a component and its possible runtime transformations, expressed by the various management controllers of the component system.

Our approach is bi-directional: from models to program, or back. We use techniques of static analysis and abstract interpretation to extract models from the code of distributed applications, or from dedicated specification formalisms [3]. On the other hand, we generate “safe by construction” code skeletons, from high level specifications; this guarantees the behavioural properties of the components. We then use generic tools from the verification community to check properties of these models. We concentrate on behavioural properties, expressed in terms of temporal logics (safety, liveness), of adequacy of an implementation to its specification and of correct composition of software components.

PHOENIX Project-Team

3. Scientific Foundations

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.,* [48]), providing some level of abstraction above programming, or (2) integrated into a programming language (*e.g.,* [20], [51]), 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 the Design of Software Systems

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 [31]. 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. [1] 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 [9], access conflicts to resources [34], and quality of service constraints [32].

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.

[1] The Error Model Annex is a standardized AADL extension for the description of errors [52].

PLANETE Project-Team

3. Scientific Foundations

3.1. Experimental approach to Networking

Based on a practical view, the Planète approach to address the above research topics is to design new communication protocols or mechanisms, to implement and to evaluate them either by simulation or by experimentation on real network platforms (such as PlanetLab and OneLab). Our work includes a substantial technological component since we implement our mechanisms in pre-operational systems and we also develop applications that integrate the designed mechanisms as experimentation and demonstration tools. We also work on the design and development of networking experimentation tools such as the ns-3 network simulator and experimental platforms. We work in close collaboration with research and development industrial teams.

In addition to our experimentation and deployment specificities, we closely work with researchers from various domains to broaden the range of techniques we can apply to networks. In particular, we apply techniques of the information and queuing theories to evaluate the performance of protocols and systems. The collaboration with physicists and mathematicians is, from our point of view, a promising approach to find solutions that will build the future of the Internet.

In order to carry out our approach as well as possible, it is important to attend and contribute to IETF (Internet Engineering Task Force) and other standardization bodies meetings on a regular basis, in order to propose and discuss our ideas in the working groups related to our topics of interests.

RAP Project-Team

3. Scientific Foundations

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 content-centric 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. Scientific Foundations

3.1. Research rationale

Peer-to-peer, Cloud computing, distributed system, data consistency, fault tolerance, dynamic adaptation, large-scale environments, replication.

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.[1]

While our work historically focused on distributed systems, we now cover a larger part of the whole Computer Systems spectrum. Our topics now also include Managed Run-time Environments (MREs, a.k.a. language-level virtual machines) and operating system kernels. 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 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 peer-to-peer (P2P) and overlay networks, dynamic wireless networks, cloud computing, and manycore machines. The scientific challenges are scalability, fault tolerance, dynamicity and virtualization of physical infrastructure. Algorithms designed for classical distributed systems, such as resource allocation, data storage and placement, and concurrent 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.

[1] From the web page of ACM Transactions on Computer Systems: *"The term* 'computer systems' *is interpreted broadly and includes systems architectures, operating systems, distributed systems, and computer networks."* See http://tocs.acm.org/.

Recent multicore architectures are highly diverse. Compiling and optimizing programs for such architectures can only be done for a given target. In this setting, MREs are an elegant approach since they permit distributing a unique binary representation of an application, to which architecture-specific optimizations can be applied late on the execution machine. Finally, the concurrency provided by multicore architectures also induces new challenges for software robustness. We consider this problem in the context of systems software, using static analysis of the source code and the technology developed in the Coccinelle tool.

RMOD Project-Team

3. Scientific Foundations

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 [37], [36]. 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 develop a set of conceptual tools to support this understanding.

Some approaches based on Formal Concept Analysis (FCA) [65] 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 [23]. 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 [53]. Until now, few works have attempted to identify layers in practice: Mudpie [67] is a first cut at identifying cycles between packages as well as package groups potentially representing layers. DSM (dependency structure matrix) [66], [61] 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 [49]. 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 [68], [40].

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 apparition 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 [63], [38] and classboxes [24] 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 [63], [38]. 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 [35], [57], [25], [39] and several type systems were defined [41], [64], [58], [51].

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 [27]. 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 [38], stateful [26], and freezable [39]) 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 [56]. 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 [33] then from Smalltalk [44].

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 [48]). 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 [29], as well as controlling the access to reflective features [30], [31] are important challenges. We plan to:

- Study the isolation abstractions available in erights (http://www.erights.org) [55], [54], and Java's class loader strategies [50], [45].
- Categorize the different reflective features of languages such as CLOS [47], Python and Smalltalk [59] and identify suitable isolation mechanisms and infrastructure [42].
- Assess different isolation models (access rights, capabilities [60]...) 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 [29],

- the use encapsulation policies as a basis to restrict the interfaces of the controlled objects [62],
- 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...) [59].

ROMA Team (section vide)

RUNTIME Project-Team

3. Scientific Foundations

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.* [36], 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. In the ANR project PetaQCD, 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 are submitted to publication.

For parallel programs running on multicores, measuring reliable performance and determining performance stability is becoming a key issue: indeed, a number of hardware mechanisms may cause performance instability from one run to the other. Thread migration, memory contention (on any level of the cache hierarchy), scheduling policy of the runtime can introduce some variation, indenpendently of the program input. A speed-up is interesting only it corresponds to a performance that can be obtained through repeated execution of the application. Very few research efforts have been made in the identification of program optimization/runtime policy/hardware mechanisms that may introduce performance instability. We studied in [37] on a large set of OpenMP benchmarks performance variations, identified the mechanisms causing them and showing the need for better strategies for measuring speed-ups. Following this effort, we developped inside the tool MAQAO (Modular Assembler Quality Analyzer and Optimizer), the precise analysis of the interactions between OpenMP threads, through static analysis of binary codes and memory tracing. In particular, the influence of thread affinity is estimated and the tool proposes hints to the user to improve its OpenMP codes.

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.

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. How-

ever 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.

SARDES Project-Team

3. Scientific Foundations

3.1. Components and semantics

The primary foundations of the software component technology developed by SARDES relate to the component-based software engineering [92], and software architecture [90] fields. Nowadays, it is generally recognized that component-based software engineering and software architecture approaches are crucial to the development, deployment, management and maintenance of large, dependable software systems [41]. Several component models and associated architecture description languages have been devised over the past fifteen years: see e.g. [71] for an analysis of recent component models, and [75], [47] for surveys of architecture description languages.

To natively support configurability and adaptability in systems, SARDES component technology also draws from ideas in reflective languages [66], and reflective middleware [69], [45], [52]. Reflection can be used both to increase the separation of concerns in a system architecture, as pioneered by aspect-oriented programming [67], and to provide systematic means for modifying a system implementation.

The semantical foundations of component-based and reflective systems are not yet firmly established, however. Despite much work on formal foundations for component-based systems [72], [36], several questions remain open. For instance, notions of program equivalence when dealing with dynamically configurable capabilities, are far from being understood. To study the formal foundations of component-based technology, we try to model relevant constructs and capabilities in a process calculus, that is simple enough to formally analyze and reason about. This approach has been used successfully for the analysis of concurrency with the π-calculus [78], or the analysis of object-orientation [37]. Relevant developments for SARDES endeavours include behavioral theory and coinductive proof techniques [87], [85], process calculi with localities [48], [50], [53], and higher-order variants of the π-calculus [86], [60].

3.2. Open programming

Part of the language developments in SARDES concern the challenge of providing programming support for computer systems with continuously running services and applications, that operate at multiple physical and logical locations, that are constantly introduced, deployed, and combined, that interact, fail and evolve all the time. Programming such systems – called *open programming* by the designers of the Alice programming language [83] — is challenging because it requires the combination of several features, notably: (i) *modularity*, i.e. the ability to build systems by combining and composing multiple elements; (ii) *security*, i.e. the ability to deal with unknown and untrusted system elements, and to enforce if necessary their isolation from the rest of the system; (iii) *distribution*, i.e. the ability to build systems out of multiple elements executing separately on multiple interconnected machines, which operate at different speed and under different capacity constraints, and which may fail independently; (iv) *concurrency*, i.e. the ability to deal with multiple concurrent events, and non-sequential tasks; and (v) *dynamicity*, i.e. the ability to introduce new systems, as well as to remove, update and modify existing ones, possibly during their execution.

The rigorous study of programming features relate to the study of programming language constructs and semantics [79], [94], in general. Each of the features mentioned above has been, and continues to be, the subject of active research on its own. Combining them into a practical programming language with a well-defined formal semantics, however, is still an open question. Recent languages that provide relevant background for SARDES' research are:

- For their support of dynamic notions of modules and software components: Acute [88], Alice [83], [84], ArchJava [38], Classages [73], Erlang [40], Oz [94], and Scala [80].
- For their security and failure management features: Acute, E [77], Erlang and Oz [51].
- For their support for concurrent and distributed execution, Acute, Alice, JoCaml [56], E, Erlang, Klaim [44], and Oz.

3.3. Software infrastructure

The SARDES approach to software infrastructure is both architecture-based and language-based: architecture-based for it relies on an explicit component structure for runtime reconfiguration, and language-based for it relies on a high-level type safe programming language as a basis for operating system and middleware construction. Exploiting high-level programming languages for operating system construction [91] has a long history, with systems such as Oberon [95], SPIN [43] or JX [57]. More recent and relevant developments for SARDES are:

- The developments around the Singularity project at Microsoft Research [55], [63], which illustrates the use of language-based software isolation for building a secure operating system kernel.
- The seL4 project [58], [68], which developed a formal verification of a modern operating system microkernel using the Isabelle/HOL theorem prover.
- The development of operating system kernels for multicore hardware architectures such as Corey [46] and Barrelfish [42].
- The development of efficient run-time for event-based programming on multicore systems such as libasync [96], [70].

3.4. System management and control

Management (or *Administration*) is the function that aims at maintaining a system's ability to provide its specified services, with a prescribed quality of service. We approach management as a *control* activity, involving an event-reaction loop: the management system detects events that may alter the ability of the managed system to perform its function, and reacts to these events by trying to restore this ability. The operations performed under system and application administration include observation and monitoring, configuration and deployment, resource management, performance management, and fault management.

Up to now, administration tasks have mainly been performed in an ad-hoc fashion. A great deal of the knowledge needed for administration tasks is not formalized and is part of the administrators' know-how and experience. As the size and complexity of the systems and applications are increasing, the costs related to administration are taking up a major part of the total information processing budgets, and the difficulty of the administration tasks tends to approach the limits of the administrators' skills. For example, an analysis of the causes of failures of Internet services [81] shows that most of the service's downtime may be attributed to management errors (e.g. wrong configuration), and that software failures come second. In the same vein, unexpected variations of the load are difficult to manage, since they require short reaction times, which human administrators are not able to achieve.

The above motivates a new approach, in which a significant part of management-related functions is performed automatically, with minimal human intervention. This is the goal of the so-called *autonomic computing* movement [64]. Several research projects [35] are active in this area. [65], [62] are recent surveys of the main research problems related to autonomic computing. Of particular importance for SARDES' work are the issues associated with configuration, deployment and reconfiguration [54], and techniques for constructing control algorithms in the decision stage of administration feedback loops, including discrete control techniques [49], and continuous ones [59].

Management and control functions built by SARDES require also the development of distributed algorithms [74], [93] at different scales, from algorithms for multiprocessor architectures [61] to algorithms for cloud computing [76] and for dynamic peer-to-peer computing systems [39], [82]. Of particular relevance in the latter contexts are epidemic protocols such as gossip protocols [89] because of their natural resilience to node dynamicity or *churn*, an inherent scalability.

SCORE Team

3. Scientific Foundations

3.1. Introduction

Our scientific foundations are grounded on distributed collaborative systems supported by sophisticated data sharing mechanisms and an 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 Score 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 Score team investigated the issues related to the management of distributed shared data and coordination between users and groups.

Service oriented Computing [29] is an established domain on which the ECOO and now the Score team has 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 coordinate 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 spans the disciplines of distributed computing, software engineering and 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) [32] 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 [31] 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 [27]
- the algorithms based on commutative replicated data types (CRDT) [30]

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 [28] 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. Business Process Management

Business Process Management (BPM) is considered as a core discipline behind Service Management and Computing. BPM, that includes the analysis, the modelling, the execution, the monitoring and the continuous improvement of enterprise processes is for us a central domain of studies.

A lot of efforts has been devoted in the past years to established standards business process models founded on well grounded theories (e.g. Petri Nets) that meet the needs of both business analyst 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 artifacts. Part of our work has been an attempt to reconcile these point of views, leading on one side to the 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 application providing value based on the composition of lower level services [26].

3.5. Service Composition

More and more, we are considering processes as piece 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 [25].

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 previous results an event oriented composition framework DISC, we have obtain new results dedicated to the runtime verification of violations in services choreographies [6], [7], [12]

SOCRATE Team

3. Scientific Foundations

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 axis 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 physical protocols layer on these software radio machines. Figure 3 illustrates the three region of a transceiver corresponding to the three Socrate axes.

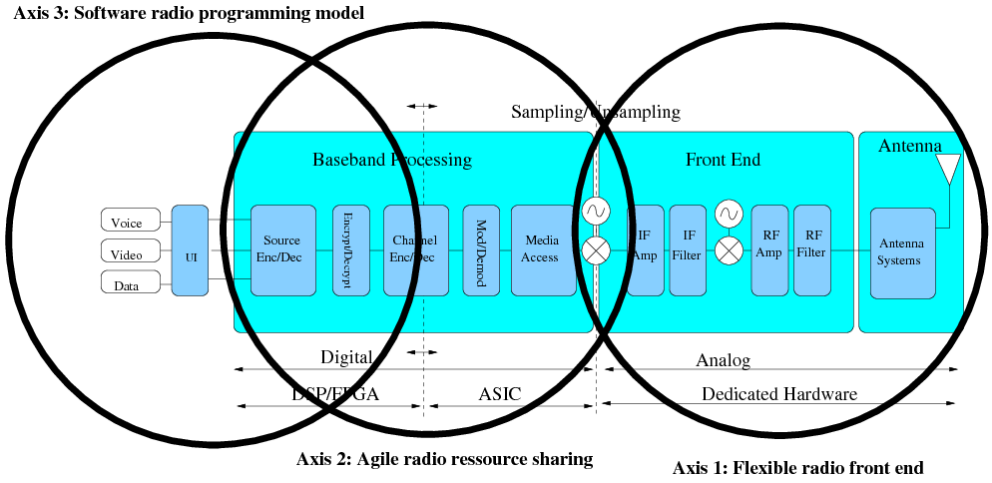


Figure 3. Center of interest for each ot the three Socrate researche axes with respect to a generic software radio terminal.

3.2. Flexible Radio Front-End

Guillaume Villemaud (coordinator), 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 amplifier which are currently studied by hardware oriented teams. The purpose of this axe is to build system level simulation taking into account the state of the art of each key components. A large part of this work will be supported in the frame of the FUI project EconHome starting in January 2011.

3.3. Agile Radio Resource Sharing

Jean-Marie Gorce (coordinator), Claire Goursaud, Nikolai Lebedev

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 agiles : 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 about 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 MANETS, 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.

3.4. Software Radio Programming Model

Tanguy Risset (coordinator), Kevin Marquet, Guillaume Salagnac

Finally the third research axis is concerned with software aspect of the software radio terminal (left of Fig 3). We have currently two action 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[1].
- 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

[1] Lyrtech (http://www.lyrtech.com) designs and sells radio card receivers with multiple antennas offering the possibility to implement a complete communication stack

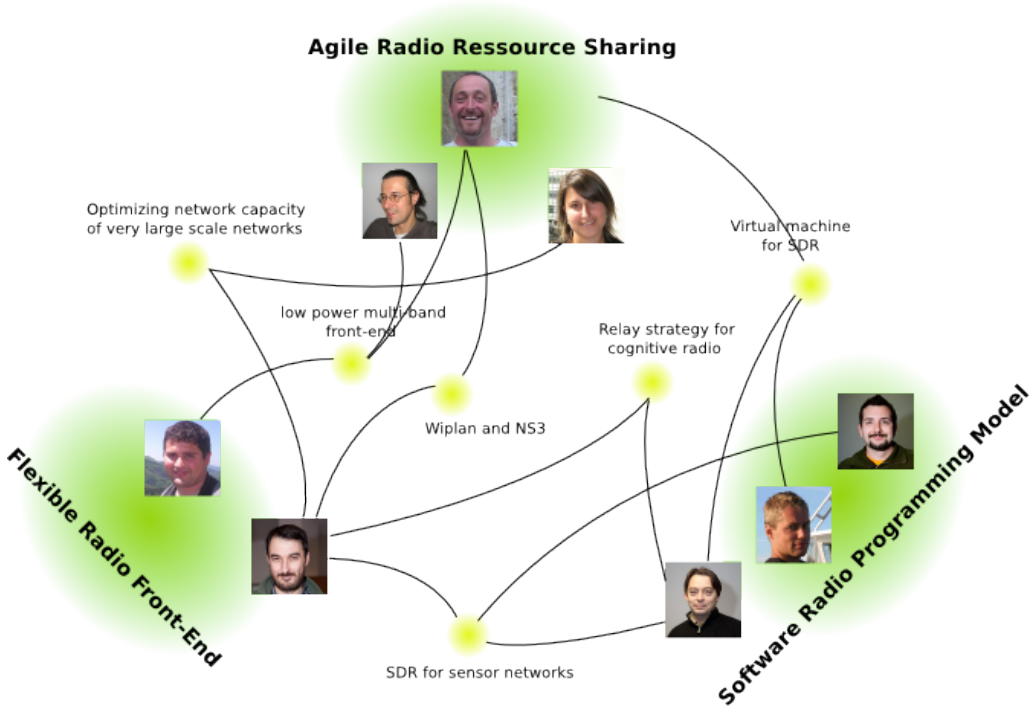


Figure 4. Inter-Axis Collaboration in Socrate: we expect innovative results to come from this pluri-disciplinary research

As mentioned earlier, innovative results will come from collaborations between these 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 axis, 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) are planned to collaborate thanks to the complementarity of their subjects.
- *SDR for sensor networks*. A master student have been hired in 2012 and a PhD should be started in collaboration with FT R&D, involving people from axis 3 (Guillaume Salagnac, Tanguy Risset) and axis 1 (Guillaume Villemaud).
- *Wiplan and NS3*. The MobiSim ADT and iPlan projects involve 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).
- *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 where 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.

TREC Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

- **Modeling and performance analysis of wireless networks.** Our main focus was on cellular networks, mobile ad hoc networks (MANETs) and their vehicular variants called VANETs.

  Our main advances about wireless networks have been based on the development of analytical tools for their performance analysis and on new results from network information theory.

  Concerning cellular networks, the main questions bear on coverage and capacity in large CDMA networks when taking intercell interferences and power control into account. Our main focus has been on the design of: 1) a strategy for the densification and parameterization of UMTS and future OFDM networks that is optimized for both voice and data traffic; 2) new self organization and self optimization protocols for cellular networks e.g. for power control, sub-carrier selection, load balancing, etc.

  Concerning MANETs, we investigated MAC layer scheduling algorithms, routing algorithms and power control. The MAC protocols we considered are based on Aloha and CSMA as well as their cognitive radio extensions. We investigated opportunistic routing schemes for MANETs and VANETs. The focus was on cross layer optimizations allowing one to maximize the transport capacity of multihop networks.
- **Theory of network dynamics.** TREC is pursuing the analysis of network dynamics by algebraic methods. The mathematical tools are those of discrete event dynamical systems: semi-rings, and in particular network calculus, ergodic theory, perfect simulation, stochastic comparison, inverse problems, large deviations, etc. Network calculus gives results on worst-case performance evaluation; ergodic theory is used to assess the stability of discrete event dynamical systems; inverse problem methods are used to estimate some network parameters from external observations and to design network probing strategies.
- **The development of stochastic geometry and random geometric graphs tools.** Stochastic geometry 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 and also to random geometric graphs. Our research is centered on the development of a methodology for the analysis, the synthesis, the optimization and the comparison of architectures and protocols to be used in wireless communication networks. The main strength of this method is its capacity for taking into account the specific properties of wireless links, as well as the fundamental question of scalability.
- **Combinatorial optimization and analysis of algorithms.** In this research direction started in 2007, we build upon our expertise on random trees and graphs and our collaboration with D. Aldous in Berkeley. Sparse graph structures have proved useful in a number of applications from information processing tasks to the modeling of social networks. We obtained new results in this research direction: computation of the asymptotic for the rank of the adjacency matrix of random graphs, computation of the matching number and the b-matching number of large graphs. We also applied our result to design bipartite graph structures for efficient balancing of heterogeneous loads and to analyze the flooding time in random graphs.
- **Economics of networks** The premise of this relatively new direction of research, developed jointly with Jean Bolot [SPRINT ATL and then TECHNICOLOR] is that economic incentives drive the development and deployment of technology. Such incentives exist if there is a market where suppliers and buyers can meet. In today's Internet, such a market is missing. We started by looking at the general problem of security on Internet from an economic perspective.

TRISKELL Project-Team

3. Scientific Foundations

3.1. Model Driven Engineering for Distributed Software

Objects, design patterns, software components, contracts, aspects, models, UML, product lines

3.1.1. Software Product Lines

It is seldom the case nowadays that we can any longer deliver software systems with the assumption that one-size-fits-all. We have to handle many variants accounting not only for differences in product functionalities (range of products to be marketed at different prices), but also for differences in hardware (e.g.; graphic cards, display capacities, input devices), operating systems, localization, user preferences for GUI ("skins"). Obvioulsy, we do not want to develop from scratch and independantly all of the variants the marketing department wants. Furthermore, all of these variant may have many successive versions, leading to a two-dimensional vision of product-lines.

3.1.2. Object-Oriented Software Engineering

The object-oriented approach is now widespread for the analysis, the design, and the implementation of software systems. Rooted in the idea of modeling (through its origin in Simula), object-oriented analysis, design and implementation takes into account the incremental, iterative and evolutive nature of software development [76], [71]: large software system are seldom developed from scratch, and maintenance activities represent a large share of the overall development effort.

In the object-oriented standard approach, objects are instances of classes. A class encapsulates a single abstraction in a modular way. A class is both *closed*, in the sense that it can be readily instanciated and used by clients objects, and *open*, that is subject to extensions through inheritance [79].

3.1.3. Design Pattern

Since by definition objects are simple to design and understand, complexity in an object-oriented system is well known to be in the *collaboration* between objects, and large systems cannot be understood at the level of classes and objects. Still these complex collaborations are made of recurring patterns, called design patterns. The idea of systematically identifying and documenting design patterns as autonomous entities was born in the late 80's. It was brought into the mainstream by such people as Beck, Ward, Coplien, Booch, Kerth, Johnson, etc. (known as the Hillside Group). However the main event in this emerging field was the publication, in 1995, of the book *Design Patterns: Elements of Reusable Object Oriented Software* by the so-called Gang of Four (GoF), that is E. Gamma, R. Helm, R. Johnson and J. Vlissides [75]. Today, design patterns are widely accepted as useful tools for guiding and documenting the design of object-oriented software systems. Design patterns play many roles in the development process. They provide a common vocabulary for design, they reduce system complexity by naming and defining abstractions, they constitute a base of experience for building reusable software, and they act as building blocks from which more complex designs can be built. Design patterns can be considered reusable micro-architectures that contribute to an overall system architecture. Ideally, they capture the intent behind a design by identifying the component objects, their collaborations, and the distribution of responsibilities. One of the challenges addressed in the Triskell project is to develop concepts and tools to allow their formal description and their automatic application.

3.1.4. Component

The object concept also provides the basis for *software components*, for which Szyperski's definition [86] is now generally accepted, at least in the industry:

> *A software component is a unit of composition with contractually specified interfaces and explicit context dependencies only. A software component can be deployed independently and is subject to composition by third party.*

Component based software relies on assemblies of components. Such assemblies rely in turn on fundamental mechanisms such as precise definitions of the mutual responsability of partner components, interaction means between components and their non-component environment and runtime support (e.g. .Net, EJB, Corba Component Model CCM, OSGI or Fractal).

Components help reducing costs by allowing reuse of application frameworks and components instead of redeveloping applications from scratch (product line approach). But more important, components offer the possibility to radically change the behaviors and services offered by an application by substitution or addition of new components, even a long time after deployment. This has a major impact of software lifecycle, which should now handle activities such as the design of component frameworks, the design of reusable components as deployment units, the validation of component compositions coming from various origins and the component life-cycle management.

Empirical methods without real component composition models have appeared during the emergence of a real component industry (at least in the Windows world). These methods are now clearly the cause of untractable validation and of integration problems that can not be transposed to more critical systems (see for example the accidental destruction of Ariane 501 [78]).

Providing solutions for formal component composition models and for verifiable quality (notion of *trusted components*) are especially relevant challenges. Also the methodological impact of component-based development (for example within the maturity model defined by the SEI) is also worth attention.

3.1.5. Contracts

Central to this trusted component notion is the idea of *contract*. A software contract captures mutual requirements and benefits among stake-holder components, for example between the client of a service and its suppliers (including subcomponents). Contracts strengthen and deepen interface specifications. Along the lines of abstract data type theory, a common way of specifying software contracts is to use boolean assertions called pre- and post-conditions for each service offered, as well as class invariants for defining general consistency properties. Then the contract reads as follows: The client should only ask a supplier for a service in a state where the class invariant and the precondition of the service are respected. In return, the supplier promises that the work specified in the post-condition will be done, and the class invariant is still respected. In this way rights and obligations of both client and supplier are clearly delineated, along with their responsibilities. This idea was first implemented in the Eiffel language [80] under the name *Design by Contract*, and is now available with a range of expressive power into several other programming languages (such as Java) and even in the Unified Modeling Language (UML) with the Object Constraint Language (OCL) [87]. However, the classical predicate based contracts are not enough to describe the requirements of modern applications. Those applications are distributed, interactive and they rely on resources with random quality of service. We have shown that classical contracts can be extended to take care of synchronization and extrafunctional properties of services (such as throughput, delays, etc) [69].

3.1.6. Models and Aspects

As in other sciences, we are increasingly resorting to modelling to master the complexity of modern software development. According to Jeff Rothenberg,

> *Modeling, in the broadest sense, is the cost-effective use of something in place of something else for some cognitive purpose. It allows us to use something that is simpler, safer or cheaper than reality instead of reality for some purpose. A model represents reality for the given purpose; the model is an abstraction of reality in the sense that it cannot represent all aspects of reality. This allows us to deal with the world in a simplified manner, avoiding the complexity, danger and irreversibility of reality.*

So modeling is not just about expressing a solution at a higher abstraction level than code. This has been useful in the past (assembly languages abstracting away from machine code, 3GL abstracting over assembly languages, etc.) and it is still useful today to get a holistic view on a large C++ program. But modeling goes well beyond that.

Modeling is indeed one of the touchstone of any scientific activity (along with validating models with respect to experiments carried out in the real world). Note by the way that the specificity of engineering is that engineers build models of artefacts that usually do not exist yet (with the ultimate goal of building them).

In engineering, one wants to break down a complex system into as many models as needed in order to address all the relevant concerns in such a way that they become understandable enough. These models may be expressed with a general purpose modeling language such as the Unified Modeling Language (UML), or with Domain Specific Languages when it is more appropriate.

Each of these models can be seen as the abstraction of an aspect of reality for handling a given concern. The provision of effective means for handling such concerns makes it possible to establish critical trade-offs early on in the software life cycle, and to effectively manage variation points in the case of product-lines.

Note that in the Aspect Oriented Programming community, the notion of aspect is defined in a sligthly more restricted way as the modularization of a cross-cutting concern. If we indeed have an already existing “main” decomposition paradigm (such as object orientation), there are many classes of concerns for which clear allocation into modules is not possible (hence the name “cross-cutting”). Examples include both allocating responsibility for providing certain kinds of functionality (such as loggin) in a cohesive, loosely coupled fashion, as well as handling many non-functional requirements that are inherently cross-cutting e.g.; security, mobility, availability, distribution, resource management and real-time constraints.

However now that aspects become also popular outside of the mere programming world [84], there is a growing acceptance for a wider definition where an aspect is a concern that can be modularized. The motivation of these efforts is the systematic identification, modularization, representation, and composition of these concerns, with the ultimate goal of improving our ability to reason about the problem domain and the corresponding solution, reducing the size of software model and application code, development costs and maintenance time.

3.1.7. Design and Aspect Weaving

So really modeling is the activity of separating concerns in the problem domain, an activity also called *analysis*. If solutions to these concerns can be described as aspects, the design process can then be characterized as a weaving of these aspects into a detailed design model (also called the solution space). This is not new: this is actually what designers have been effectively doing forever. Most often however, the various aspects are not *explicit*, or when there are, it is in the form of informal descriptions. So the task of the designer is to do the weaving in her head more or less at once, and then produce the resulting detailled design as a big tangled program (even if one decomposition paradigm, such as functional or object-oriented, is used). While it works pretty well for small problems, it can become a major headache for bigger ones.

Note that the real challenge here is not on how to design the system to take a particular aspect into account: there is a huge design know-how in industry for that, often captured in the form of Design Patterns (see above). Taking into account more than one aspect as the same time is a little bit more tricky, but many large scale successful projects in industry are there to show us that engineers do ultimately manage to sort it out.

The real challenge in a product-line context is that the engineer wants to be able to change her mind on which version of which variant of any particular aspect she wants in the system. And she wants to do it cheaply, quickly and safely. For that, redoing by hand the tedious weaving of every aspect is not an option.

3.1.8. Model Driven Engineering

Usually in science, a model has a different nature that the thing it models ("do not take the map for the reality" as Sun Tse put it many centuries ago). Only in software and in linguistics a model has the same nature as the thing it models. In software at least, this opens the possibility to automatically derive software from its

model. This property is well known from any compiler writer (and others), but it was recently made quite popular with an OMG initiative called the Model Driven Architecture (MDA). 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) manipulating these models to produce a detailled design that can ultimately be transformed to code or at least test suites.

The OMG has built a meta-data management framework to support the MDA. It is mainly based on a unique M3 "meta-meta-model" called the Meta-Object Facility (MOF) and a library of M2 meta-models, such as the UML (or SPEM for software process engineering), in which the user can base his M1 model.

The MDA core idea is that it should be possible to capitalize on platform-independent models (PIM), and more or less automatically derive platform-specific models (PSM) –and ultimately code– from PIM through model transformations. But in some business areas involving fault-tolerant, distributed real-time computations, there is a growing concern that the added value of a company not only lies in its know-how of the business domain (the PIM) but also in the design know-how needed to make these systems work in the field (the transformation to go from PIM to PSM). Reasons making it complex to go from a simple and stable business model to a complex implementation include:

- Various modeling languages used beyond UML,
- As many points of views as stakeholders,
- Deliver software for (many) variants of a platform,
- Heterogeneity is the rule,
- Reuse technical solutions across large product lines (e.g. fault tolerance, security, etc.),
- Customize generic transformations,
- Compose reusable transformations,
- Evolve and maintain transformations for 15+ years.

This wider context is now known as Model Driven Engineering.

URBANET Team

3. Scientific Foundations

3.1. Capillary networks

The digital cities that evolve today need a thin and dense digitalization of their citizens and infrastructures' activities. There is hence a need for a new networking paradigm capable of providing enough capacity and quality of service. From the user point of view, there is only one network to access to data and applications, but from the point of view of operators, engineers, there are several access networks: wireless sensor networks to measure the physical world, the cellular networks (including 3G/4G) to handle mobility, mesh networks to support new applications and services.

We propose to aggregate all these networks in the concept of capillary network. A capillary network is, for the user or end device, a link to Internet, whatever the link is. For engineers and researchers, a capillary network represents all the different possible paths we have from the user terminal to the access network. Providing the support for a digital city and for a digital society requires to focus on Capillary Networking issues. These issues include classical challenges related to sensor, mesh, or user-centric networks (such as cellular or vehicular networks), but also present important components generated by the urban environment.

3.2. Characterizing urban networks

A typical urban capillary network will involve a set of different communication technologies like 3G/ LTE, IEEE 802.11, WSN, inter vehicular communications and many others. Each technology relies on a set of mechanisms that were designed to provide a dedicated set of functionalities. Typical mechanisms include resource allocation, scheduling, error detection and correction, routing etc.

Dimensioning the operating parameters of such network mechanisms in order to provide the desired services while ensuring the network efficiency is a classical and yet a difficult issue. There are many directions to address this problem. For instance, one can refer to the network dimensioning and traffic-engineering approaches. Cross layer optimization and Self-organizing networks (SON) paradigm in 3G/LTE are also other perspectives to tackle this issue. However, given the complexity of the problem, most of the efforts concentrate on the mono-technological and/or the mono-service cases.

In the urban scenario, the heterogeneity of the technologies and the particularity of the urban services bring up new network-dimensioning challenges. The optimization has to be extended to the inter-technological perspective and to the multi-services standpoint. The different technologies that compose the capillary network have to inter-operate in a seamless and optimal way so that they can provide user-centric services with the desired quality of experience. Consider, for instance, dimensioning the scheduling mechanism of a mesh network, which has to carry the traffic generated by different WSN in the city. Predicting the time and spatial distribution of the traffic generated by the different WSNs are clearly among the key elements that shall be considered. On the other side, from a downlink standpoint, consider the judicious setting of an WSN aggregation mechanism accordingly with the time varying capacity of the mesh backbone level.

It is quite clear that these questions cannot be addressed without characterizing the features of an urban capillary network. This covers the geographical properties of the networks (distribution, density, nodes degree, mobility etc.) as well as the data traffic characteristics of urban services. Understanding these proprieties and their correlation is still an uncovered area. The main challenge in this case is the production of quantitative traces from real or realistic urban mobility, networks and services. For example, in urban mobility scenarios, how long devices are in radio range of each other gives temporal constraints on the communications protocols that should be understood. In this duration, devices have to self-organize or to hang on the exiting organization and to exchange information.

A second step is to derive analytical or simulation models that will be used for network dimensioning and optimization. Many models already exist in the literature in related scientific fields and they could be considered or adapted to this purpose. This covers different models ranging from radio propagation, vehicular or pedestrian mobility, traffic pattern, etc, the difficulty being on how to mix these models and how to choose the right time magnitude and spatial scale in order to preserve the accuracy of the capillary network features while maintaining the model complexity tractable. The derived models could serve to optimize the different mechanisms involved in the urban capillary network.

The inference between different networks and services is quite complex to understand and to model, therefore a simple approach would be to decouple the models. Choosing the right decoupling technique depends on the targeted temporal and spatial level of the input and output parameters. Again, the latter shall capture for each decoupled model a selected set of significant features of the capillary network. Finally, the purpose of the constructed models is to obtain the optimal dimensioning of the network mechanisms. Several optimization techniques, from exact to heuristics ones, shall be considered to compute the best operating parameters. One of the main challenges here is to maintain the computational complexity tractable by exploiting the specific structure of the problems induced by the city.

3.3. Highly scalable protocols

The networks formed in an urban environment can sometimes be particularly challenging for the MAC layer protocols and QoS support, especially if the network is not centralized or synchronized: very high node degree, unstable and asymmetric links, etc.

MAC layer protocols are either very difficult to implement in distributed and self-organized environment or present serious scaling issues. Studies focusing on distributed TDMA showed that MAC protocols from this class can be successfully designed to accommodate channel access for a high number of contending nodes. However, scalability is always obtained following a learning phase with relatively high convergence time. This means that in a dynamic network scenario like the one encountered in most urban capillary networks, the MAC protocol spends most of the time in the learning phase, where it achieves a reduced performance. The same problem appears when trying to distribute other usually centralized schemes, such as OFDMA or CDMA. On the other hand, CSMA/CA protocols are distributed by their nature.

However, the current leading solutions in this area are based on the IEEE 802.11 Distributed Coordination Function (DCF), a channel access method designed and optimized for Wireless LANs with a central access point and a maximum of 10-20 contending stations. The DCF is well-known for its scalability issues, especially in multi-hop dynamic networks, and adding energy constraints usually existing in wireless sensor networks does not improve its performance. While multiple MAC layer congestion control solutions have been proposed in the context of mobile ad-hoc networks, the approach is usually based on the idea of reducing the number of neighbors, either through transmission power control or data rate adjustment. However, this is just a workaround and the search for a truly scalable MAC layer protocol for high density wireless networks is still open.

Regarding the network layers, in order to have multi-service platforms deployed in practice, all the requirements of telecommunication operators should be present, in particular in wireless sensor and actuators networks, within the key notion of Service Level Agreement (SLA) for traffic differentiation, quality of service support (delay, reliability, etc.). Moreover, because the world becomes more and more connected to Internet, IP should be supported in wireless sensor networks. The IETF proposes the use of RPL (Routing Protocol for low power and lossy networks), where it is clear that the support of several Destination oriented Directed Acyclic Graphs (DoDAG) is required, and a complete traffic management is needed. Moreover, RPL assumes a static topology but the classical sensor networks give way to urban sensing, where the user's smartphone give the physical measures to the operators. Therefore, the data collection becomes distributed, sometimes local, the network is now dynamic. In such a scenario, inconsistencies stemming from data collected using different calibration process raise a lot of interests. Moreover, data aggregation and data gathering is, in capillary networks, at the heart of the issues related to the limited capacity of the networks. In particular, combining local aggregation and measurement redundancy for improving data reliability is a promising approach.

3.4. 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, especially the one related to future machine-to-machine communications. Cumulated with the lack of available new radio frequency spectrum, this leads to an important challenge for mobile operators, who are looking at both licensed and unlicensed technologies for solutions.

Several approaches can be taken to tackle this problem, the most obvious being to exploit the multitude of alternative network interfaces in order to prevent data to go through the cellular network. In this perspective, taking advantage of the fact that cellular operators usually possess an important ADSL or cable infrastructure for wired services, the development of femtocell solutions has become 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 is to use IEEE 802.11 Wi-Fi (or other multi-hop 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, concluding that more than 65% of the data can be offloaded from the cellular infrastructure in high density areas. 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 hot-spot connectivity and multi-hop communications is an appealing answer to broadcasting geolocalized informations efficiently.

A complementary approach, more operator oriented, for minimizing the transmission power of cellular networks as well as increasing the network capacity, consists in a dramatic increase in the deployment of micro-cells. On the other hand, increasing the number of micro-cells multiplies the energy consumed by the cells whatever their state, idle, transmitting or receiving, which is a major and growing part of the access network energy consumption. For a sustainable deployment of such micro-cell infrastructures and for a significant 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, which can be done by mechanisms inspired by the abundant works on WSNs and adapted to the energy models of micro-cells, and to the requirements of a cellular network, in particular the need for providing an adequate quality of service to dynamic and mobile clients.

ALICE Project-Team

3. Scientific Foundations

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 [25]. 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

Participants: Laurent Alonso, Dobrina Boltcheva, Alejandro Galindo, Phuong Ho, Samuel Hornus, Thomas Jost, Bruno Lévy, David Lopez, Romain Merland, Vincent Nivoliers, Jeanne Pellerin, Nicolas Ray, Dmitry Sokolov, Rhaleb Zayer.

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., [26]). 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 [5]. 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 [6]. 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, we studied last year several ways of discretizing partial differential equations on meshes, including Finite Element Modeling and Discrete Exterior Calculus. This year, we also explored Spectral Geometry Processing and Sampling Theory (more on this below).

3.3. Computer Graphics

Participants: Sylvain Lefebvre, Samuel Hornus, Bruno Lévy, Vincent Nivoliers, Nicolas Ray, Dmitry Sokolov, Rhaleb Zayer.

texture synthesis, texture mapping,

Content creation is one of the major challenge in Computer Graphics. Modelling geometries 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.

In addition to the core algorithm 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 [7] or in volumes [1].

ALPAGE Project-Team

3. Scientific Foundations

3.1. From programming languages to linguistic grammars

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Pierre Boullier.

Historically, several members of Alpage were originally specialists in the domain of modeling and parsing for programming languages, and are 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 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 [58], [96], [103]) 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 [121], [117]. Meta-Grammars 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

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 [84] or automatic [91], [92] 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 [72], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [70].

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 ...) [129], [89]. 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 [85]. 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. 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] and derive the best input for syntagmatic statistical parsing [74]. Benchmarking several PCFG-based learning frameworks [11] 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) [92].
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 [70] 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 [113].
Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [65], 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. Results are sketched in section 6.4 .

3.3. Dynamic wide coverage lexical resources

Participants: Benoît Sagot, Laurence Danlos, Rosa Stern, Éric Villemonte de La Clergerie.

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 [102]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [130],[10]. At the semantic level, automatic wordnet development tools have been described [95], [123], [82], [80]. 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* [98],[8], developed within the Alexina framework, as well as a wordnet for French, the WOLF [7], the first freely available resource of the kind.

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* and the WOLF. 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 2010 or before exist for Slovak [102], Polish [104], English, Spanish [87], [86] and Persian [108], not including freely-available lexicons adapted to the Alexina framework.

3.4. Shallow processing

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Rosa Stern.

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 obviously 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. The development of such linguistic chains, such as SXPipe , is not a trivial task [6]. Moreover, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains. In that regard, less-standard pre-processings such as word clustering have led to promising results [112].

In fact, such processing chains are mostly used as such, and not only as pre-processing tools before parsing. They aim at performing the basic tasks that produce immediately usable results for many applications, such as tokenization, sentence segmentation, spelling correction, and, most importantly, named entity detection, disambiguation and resolution.

3.5. Discourse structures

Participants: Laurence Danlos, Charlotte Roze.

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 [76].

There are three main frameworks used to model discourse structures: RST, SDRT , and, more recently, D-LTAG. Inside Alpage, Laurence Danlos has introduced D-STAG (Discourse Synchronous TAGs, [77],[5]), 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. Scientific Foundations

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) [62], graphic designers such as Bertin [51] and Tufte [61], and HCI researchers in the field of Information Visualization [50].

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 [56] and Triesman's "preattentive processing" theory [60]. 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 [54]. 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 [52]. 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 [8] [59], [57], [58], [55]. 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.

AXIS Project-Team (section vide)

AYIN Team

3. Scientific Foundations

3.1. Probabilistic approaches

Following a Bayesian methodology as far as possible, probabilistic models are used within the AYIN team for three purposes: to describe the class of images to be expected from any given scene, to describe prior knowledge about the scene and to incorporate specific constraints. The models used in AYIN fall into the following two classes.

3.1.1. Markov random fields

Markov random fields were introduced to image processing in the Eighties, and were quickly applied to the full range of inverse problems in computer vision. They owe their popularity to their flexible and intuitive nature, which makes them an ideal modelling tool, and to the existence of standard and easy-to-implement algorithms for their solution. In the AYIN team, attention is focused on their use in image modelling, in particular of textures; on the development of improved prior models for segmentation; and on the lightening of the heavy computational load traditionally associated with these techniques, in particular via the study of varieties of hierarchical random fields.

3.1.2. Stochastic geometry

One of the grand challenges of computer vision and image processing is the expression and use of prior geometric information. For satellite and aerial imagery, this problem has become increasingly important as the increasing resolution of the data results in the necessity to model geometric structures hitherto invisible. One of the most promising approaches to the inclusion of this type of information is stochastic geometry, which is an important line of research 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 distribution are called ‘marked point processes’. Such processes have been recently applied to skin care problems.

3.2. Parameter estimation

One of the most important problems studied in the AYIN team is how to estimate the parameters that appear in the models. For probabilistic models, the problem is quite easily framed, but is not necessarily easy to solve, particularly in the case when it is necessary to extract simultaneously both the information of interest and the parameters from the data.

3.3. Hierarchical models

Another line of research in the AYIN team concerns development of graph-based, in particular, hierarchical models for very high resolution image analysis and classification. A specific hierarchical model recently developed in AYIN represents an image by a forest structure, where leaf nodes represent image regions at the finest level of partition, while other nodes correspond to image regions at the coarser levels of partitions. The AYIN team is interested in developing multi-feature models of image regions as an ensemble of spectral, texture, geometrical and classification features, and establishing new criteria for comparing image regions. Recent research concerns extension of hierarchical models to a temporal dimension, for analyzing multitemporal data series.

COPRIN Project-Team

3. Scientific Foundations

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

$$\forall \{x_1, x_2, ..., x_n\} \in \{X_1, X_2, ..., X_n\}, A \leq F(x_1, x_2, ..., x_n) \leq B \tag{110}$$

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], [7], [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 [11], [19]
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 COPRIN 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

COPRIN has a long-standing tradition of robotics studies, especially for closed-loop robots [4]. 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 [13], [21] such as workspace and singularity analysis, positioning accuracy [24], trajectory planning, reliability, calibration [33], 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 [28]. The methods that we develop can be used for other robotic problems, see for example the management of uncertainties in aircraft design [34], [10].

Our theoretical work must be validated through experiments that are essential for the sake of credibility. A contrario, experiments will feed theoretical work. Hence COPRIN works with partners on the development of real robots but also develops its own prototypes. We usually develop a new robot prototype every 6 years but since 2008 we have started the development of seven new robot prototypes, mostly related to assistance robotics. Furthermore 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. We intend to make a full use of such devices, especially for assistance purpose

In term of applications we have focused up to now on the development of special machines (machine-tool, ultra-high accuracy positioning device, spatial telescope). Although this activity will be pursued, we have started in 2008 a long-term move toward *service robotics*, i.e. robots that are closer to human activity. In service robotics we are interested in domotics, smart objects, rehabilitation and medical robots [8], [9], [26] and entertainment, that can be regrouped under the name of *assistance robotics* (see section 6.2.1.3). Compared to special machines for which pricing is not an issue (up to a certain point), cost is an important element for assistance robotics. While we plan to develop simple robotic systems using only standard hardware, our work will focus on a different issue: *adaptability*. We aim at providing assistance devices that are adapted to the end-user, its trajectory of life and its environment, are easy to install (because installation uncertainties are taken into account at the design stage), have a low intrusivity and are guaranteed to fulfill a set of requirements.

DAHU Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

Dahu has strong connections with the Leo project-team in Saclay.

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. Scientific Foundations

3.1. Computer assisted monitoring and diagnosis of physical systems

keywords: monitoring,diagnosis, deep model, fault model, simulation, chronicle acquisition

Our work on monitoring and diagnosis relies on model-based approaches developed by the Artificial Intelligence community since the seminal studies by R. Reiter and J. de Kleer [63], [74]. Two main approaches have been proposed then: (i) the consistency-based approach, relying on a model of the expected correct behavior ; (ii) the abductive approach which relies on a model of the failures that might affect the system, and which identifies the failures or the faulty behavior explaining the anomalous observations. See the references [21], [23] for a detailed exposition of these investigations.

Since 1990, the researchers in the field have studied dynamic system monitoring and diagnosis, in a similar way as researchers in control theory do. What characterizes the AI approach is the use of qualitative models instead of quantitative ones and the importance given to the search for the actual source/causes of the faulty behavior. Model-based diagnosis approaches rely on qualitative simulation or on causal graphs in order to look for the causes of the observed deviations. The links between the two communities have been enforced, in particular for what concerns the work about discrete events systems and hybrid systems. Used formalisms are often similar (automata, Petri nets ,...) [28], [27].

Our team focuses on monitoring and on-line diagnosis of discrete events systems and in particular on monitoring by alarm management.

Two different methods have been studied by our team in the last years:

- In the first method, the automaton used as a model is transformed off-line into an automaton adapted to diagnosis. This automaton is called a *diagnoser*. This method has first been proposed by M. Sampath and colleagues [65]. The main drawback of this approach is its centralized nature that requires to explicitly build the global model of the system, which is most of the time unrealistic. It is why we proposed a decentralized approach in [60].
- In the second method, the idea is to associate each failure that we want to detect with a *chronicle* (or a scenario), i.e. a set of observable events interlinked by time constraints. The chronicle recognition approach consists in monitoring and diagnosing dynamic systems by recognizing those chronicles on-line [43], [62], [41].

  One of our research focus is to extend the chronicle recognition methods to a distributed context. Local chronicle bases and local recognizers are used to detect and diagnose each component. However, it is important to take into account the interaction model (messages exchanged by the components). Computing a global diagnosis requires then to check the synchronisation constraints between local diagnoses.

  Another issue is the chronicle base acquisition. An expert is often needed to create the chronicle base, and that makes the creation and the maintenance of the base very expensive. That is why we are working on an automatic method to acquire the base.

Developing diagnosis methodologies is not enough, especially when on-line monitoring is required. Two related concerns must be tackled, and are the topics of current research in the team:

- The ultimate goal is usually not merely to diagnose, but to put the system back in some acceptable state after the occurrence of a fault. One of our aim is to develop self-healable systems able to self-diagnose and -repair.

- When designing a system and equipping it with diagnosis capabilities, it may be crucial to be able to check off-line that the system will behave correctly, i.e., that the system is actually 'diagnosable'. A lot of techniques have been developed in the past (see Lafortune and colleagues [64]), essentially in automata models. We extended them to cope with temporal patterns. A recent focus has been to study the self-healability of systems (ability to self-diagnose and -repair).

3.2. Machine learning and data mining

keywords: machine learning, Inductive Logic Programming (ILP), temporal data mining, temporal abstraction, data-streams

The machine learning and data mining techniques investigated in the group aim at acquiring and improving models automatically. They belong to the field of machine or artificial learning [38]. In this domain, the goal is the induction or the discovery of hidden objects characterizations from their descriptions by a set of features or attributes. For several years we investigated Inductive Logic Programming (ILP) but now we are also working on data-mining techniques.

We are especially interested in structural learning which aims at making explicit dependencies among data where such links are not known. The relational (temporal or spatial) dimension is of particular importance in applications we are dealing with, such as process monitoring in health-care, environment or telecommunications. Being strongly related to the dynamics of the observed processes, attributes related to temporal or spatial information must be treated in a special way. Additionally, we consider that the legibility of the learned results is of crucial importance as domain experts must be able to evaluate and assess these results.

The discovery of spatial patterns or temporal relations in sequences of events involve two main steps: the choice of a data representation and the choice of a learning technique.

We are mainly interested in symbolic supervised and unsupervised learning methods. Furthermore, we are investigating methods that can cope with temporal or spatial relationships in data. In the sequel, we will give some details about relational learning, relational data-mining and data streams mining.

3.2.1. Relational learning

) Relational learning, also called inductive logic programming (ILP), lies at the intersection of machine learning, logic programming and automated deduction. Relational learning aims at inducing classification or prediction rules from examples and from domain knowledge. As relational learning relies on first order logic, it provides a very expressive and powerful language for representing learning hypotheses especially those learnt from temporal data. Furthermore, domain knowledge represented in the same language can also be used. This is a very interesting feature which enables taking into account already available knowledge and avoids starting learning from scratch.

Concerning temporal data, our work is more concerned with applying relational learning rather than developing or improving the techniques. Nevertheless, as noticed by Page and Srinivasan [59], the target application domains (such as signal processing in health-care) can benefit from adapting relational learning scheme to the particular features of the application data. Therefore, relational learning makes use of constraint programming to infer numerical values efficiently [66]. Extensions, such as QSIM [49], have also been used for learning a model of the behavior of a dynamic system [44]. Precisely, we investigate how to associate temporal abstraction methods to learning and to chronicle recognition. We are also interested in constraint clause induction, particularly for managing temporal aspects. In this setting, the representation of temporal phenomena uses specific variables managed by a constraint system [61] in order to deal efficiently with the associated computations (such as the covering tests).

For environmental data, we have investigated tree structures where a set of attributes describe nodes. Our goal is to find patterns expressed as sub-trees [37] with attribute selectors associated to nodes.

3.2.2. Data mining

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 deserved a lot of interest in the last 10 years. For instance, many enhancements have been proposed to the well-known Apriori algorithm [25]. 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" [26], which extends Apriori by coping with the order of elements in patterns.

In [54], Mannila and Toivonen extended the work of Aggrawal et al. by introducing an algorithm for mining patterns involving temporal episodes with a distinction between parallel and sequential event patterns. Later, in [42], Dousson and Vu Duong introduced an algorithm for mining chronicles. Chronicles are sets of events associated with temporal constraints on their occurrences. They generalize the temporal patterns of Mannila and Toivonen. The candidate generation is an Apriori-like algorithm. The chronicle recognizer CRS [40] is used to compute the support of patterns. Then, the temporal constraints are computed as an interval whose bounds are the minimal and the maximal temporal extent of the delay separating the occurrences of two given events in the data-set. Chronicles are very interesting because they can model a system behavior with sufficient precision to compute fine diagnoses. Their extraction from a data-set is reasonably efficient. They can be efficiently recognized on an input data stream.

Relational data-mining [22] can be seen as generalizing these works to first order patterns. In this field, the work of Dehaspe for extracting first-order association rules have strong links with chronicles. Another interesting research concerns inductive databases which aim at giving a theoretical and logical framework to data-mining [50], [39]. In this view, the mining process means to query a database containing raw data as well as patterns that are implicitly coded in the data. The answer to a query is, either the solution patterns that are already present in the database, or computed by a mining algorithm, e.g., Apriori. The original work concerns sequential patterns only [53]. We have investigated an extension of inductive database where patterns are very close to chronicles [69].

3.2.3. Mining data streams

) During the last years, a new challenge has appeared in the data mining community: mining from data streams [24]. Data coming for example from monitoring systems observing patients or from telecommunication systems arrive in such huge volumes that they cannot be stored in totality for further processing: the key feature is that “you get only one look at the data” [46]. Many investigations have been made to adapt existing mining algorithms to this particular context or to propose new solutions: for example, methods for building synopses of past data in the form of or summaries have been proposed, as well as representation models taking advantage of the most recent data. Sequential pattern stream mining is still an issue [55]. At present, research topics such as, sampling, summarizing, clustering and mining data streams are actively investigated.

A major issue in data streams is to take into account the dynamics of process generating data, i.e., the underlying model is evolving and, so, the extracted patterns have to be adapted constantly. This feature, known as *concept drift* [71], [51], occurs within an evolving system when the state of some hidden system variables changes. This is the source of important challenges for data stream mining [45] because it is impossible to store all the data for off-line processing or learning. Thus, changes must be detected on-line and the current mined models must be updated on line as well.

E-MOTION Project-Team (section vide)

EXMO Project-Team

3. Scientific Foundations

3.1. Knowledge representation semantics

We usually work with semantically defined knowledge representation languages (like description logics [28], conceptual graphs and object-based languages). 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 annotation of resources within the framework of the semantic web. 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 multimedia specification languages and alignment languages, in order to establish the properties of computer manipulations of the representations.

3.2. Ontology 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. 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, that hold between these entities.

An alignment between two ontologies o and o' is a set of correspondences $\langle e, e', r \rangle$ in which:

- 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 correspodence.

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 set 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 will select the local models which 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 instanciation). Confidence measures are usually given a value between 0 and 1 and are used for expressing preferences between two correspondences.

FLOWERS Project-Team

3. Scientific Foundations

3.1. Scientific Foundations

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 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" [120]. 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 [90], 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 [97] [122]. The approach of developmental robotics consists in importing and implementing concepts and mechanisms from developmental psychology [102], cognitive linguistics [79], 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 [76] [110], grounding [88], situatedness [70], self-organization [118] [111], enaction [121], and incremental learning [77].

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 [14]. 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 [102], 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" [72] [84]. 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 [74] [80] [82]. 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 [81] [91] [116]. Based on this, a number of researchers have began in the past few years to build computational implementation of intrinsic motivation [14] [108] [114] [73] [92] [100] [115]. 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 [108][14] [109] [113]. 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 [78] [89] 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 [102]. 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 [71]. 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 [119], [75], motivated by the various mechanisms that allow humans to socially guide a robot [112]. 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 [119], [94] [101].

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 [104], [106]. We also work on developing perceptual capabilities by gradually segmenting the perceptual space and identifying objects and their characteristics through interaction with the user [48] and robots experiments [43]. 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 [96].

GRAPHIK Project-Team

3. Scientific Foundations

3.1. Logic-based Knowledge Representation and Reasoning

We follow the mainstream *logical* 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 *query answering* problem is a topical problem (see Sect. 3.3). It asks for the set of answers to a query in the KB. In the special 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 reasonings 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 is a central problem in knowledge representation and in database theory. 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 were 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 can be seen as an abstraction of ontological knowledge expressed in the main languages used in the context of KB querying. See Sect. 6.1 for details on the results obtained.

A problem generalising 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 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 Sect. 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 will 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 reasonings (see ANR project ASPIQ in Sect. 8.1), as well as argumentation and preferences (see Sect. 6.2).

IMAGINE Team

3. Scientific Foundations

3.1. A failure of standard modeling techniques?

Surprisingly, in our digital age, conceptual design of static shapes, motion and stories is almost never done on computers. Designers prefer to use traditional media even when a digital model is eventually created for setups such as industrial prototyping, and even when the elements to be designed are aimed at remaining purely virtual, such as in 3D films or games. In his keynote talk at SIGGRAPH Asia 2008, Rob Cook, vice president of technology at Pixar Animation Studios, stressed that even trained computer artists tend to avoid the use of 3D computerized tools whenever possible. They use first pen and paper, and then clay to design shapes; paper to script motion; and hand-sketched storyboards to structure narrative content and synchronise it with speech and music. Even lighting and dramatic styles are designed using 2D painting tools. The use of 3D graphics is avoided as much as possible at all of these stages, as if one could only reproduce already designed material with 3D modelling software, but not create directly with it. This disconnect can be thought of as the number one failure of digital 3D modelling methodologies. As Cook stressed: "*The new grand challenge in Computer Graphics is to make tools as transparent to the artists as special effects were made transparent to the general public*" (Cook 2008). The failure does not only affect computer artists but many users, from engineers and scientists willing to validate their ideas on virtual prototypes, to media, educators and the general public looking for simple tools to quickly personalize their favourite virtual environment.

Analyzing the reasons for this failure we observe that 3D modeling methodologies did not evolve much in the last 20 years. Standard software, such as Maya and 3dsMax, provide sophisticated interfaces to fully control all degrees of freedom and bind together an increasing number of shape and motion models. Mastering this software requires years of training to become skilled. Users have to choose the best suited representation for each individual element they need to create, and fully design a shape before being able to define its motion. In many cases, neither descriptive models, which lack high level constraints and leave the quality of results in user's hands, nor procedural ones, where realistic simulation comes at the price of control, are really convenient. A good example is modelling of garments for virtual characters. The designer may either sculpt the garment surface at rest, which provides direct control on the folds but requires lots of skill due to the lack of constraints (such as enforcing a cloth surface to be developable onto a plane), or they can tune the parameters of a physically-based model simulating cloth under gravity, which behaves as a black box and may never achieve the expected result. No mechanism is provided to roughly draft a shape, and help the user progressively improve and refine it.

Capture and reconstruction of real-world objects, using either 3D scanners or image-based methods, provides an appealing alternative for quickly creating 3D models and attracted a lot of attention from both Computer Graphics and Computer Vision research communities the last few years. Similarly, techniques for capture and reuse of real motion, enabling an easy generation of believable animation content, were widely investigated. These efforts are much welcome, since being able to embed existing objects and motion in virtual environments is extremely useful. However, it is not sufficient. One cannot scan every blade of grass, or even every expressive motion, to create a convincing virtual world. What if the content to be modelled does not exist yet, or will never exist? One of the key motivations for using digital modelling in the first place is as a tool for bringing to life new, imaginary content.

3.2. Long term vision: an "expressive virtual pen" for animated 3D content

Stepping back and taking a broader viewpoint, we observe that humans need a specialized medium or tool, such as pen and paper or a piece of clay, to convey shapes, and more generally animated scenes. Pen and paper, probably the most effective media to use, requires sketching from different viewpoints to fully represent a shape and requires a large set of drawings over time to communicate motion and stories.

Could digital modeling be turned into a tool, even more expressive and simpler to use than a pen, to quickly convey and refine shapes, motions, and stories?

This is the long term vision towards which we would like to advance.

3.3. Methodology: "Control to the user, Knowledge to the system"

Thinking of future digital modeling technologies as an "expressive virtual pen", enabling to seamlessly design, refine and convey animated 3D content, is a good source of inspiration. It led us to the following methodology:

- As when they use a pen, users should not be restricted to the editing of preset shapes or motion, but should get a **full control over their design**. This control should ideally be as easy and intuitive as when sketching, which leads to the use of gestures – although not necessarily sketching gestures – rather than of standard interfaces with menus, buttons and sliders. Ideally, these control gestures should drive the choice of the underlying geometric model, deformation tool, and animation method in a predictable but transparent way, enabling users to concentrate on their design.
- Secondly, similarly to when they draw in real, users should only have to **suggest** the 3D nature of a shape, the presence of repetitive details, or the motion or deformations that are taking place: this will allow for faster input and enable coarse to fine design, with immediate visual feedback at every stage. The modeling system should thus act similarly to a human viewer, who can imagine a 3D shape in motion from very light input such as a raw sketch. Therefore, as much as possible **a priori knowledge** should be incorporated into the models and used for inferring the missing data, leading to the use of high-level representations enabling procedural generation of content. Note that such models will also help the user towards high-quality content, since they will be able to maintain specific geometric or physical laws. Since this semi-automatic content generation should not spoil user's creativity and control, editing and refinement of the result should be allowed throughout the process.
- Lastly, creative design is indeed a matter of trial and error. We believe that creation more easily takes place when users can immediately see and play with a first version of what they have in mind, serving as support for refining their thoughts. Therefore, important features towards effective creation are to provide **real-time response** at every stage, as well as to help the user exploring the content they have created thanks to intelligent cameras and other cinematography tools.

To advance in these directions, we believe that models for shape, motion and cinematography need to be re-thought from a user centered perspective. We borrowed this concept from the Human Computer Interaction domain, but we are not referring here to **user-centred system design** (Norman 86). We rather propose to extend the concept, and develop user-centred graphical models: Ideally, a user-centred model should be designed to behave, under editing actions, the way a human user would have predicted. Editing actions may be for instance 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 used to transfer of some features from existing models to other ones. User-centred models need to incorporate knowledge in order to seamlessly generate the appropriate content from such actions. Knowledge may be for instance about developability to model paper or cloth; about constant volume to deform virtual clay or animate plausible organic shapes; about physical laws to control passive objects; or about film editing rules to generate semi-autonomous camera with planning abilities.

These user-centred models will be applied to the development of various interactive creative systems, not only for static shapes, but also for motion and stories. Although unusual, we believe that thinking about these different types of content in a similar way will enable us to improve our design principles thanks to cross fertilization between domains, and allow for more thorough experimentation and validation. The expertise we developed in our previous research team EVASION, namely the combination of layered models, adaptive degrees of freedom, and GPU computations for interactive modeling and animation, will be instrumental to ensure real-time performances. Rather than trying to create a general system that would solve everything, we plan to develop specific applications (serving as case studies), either brought by the available expertise in our

research group or by external partners. This way, user expectations should be clearly defined and final users will be available for validation. Whatever the application, we expect the use of knowledge-based, user-centred models driven by intuitive control gesture to increase both the efficiency of content creation and the quality of results.

3.4. Validation methodology

When developing digital creation tools, validation is a major challenge. Researchers working on ground-truth reconstruction can apply standard methodologies to validate their techniques, such as starting by testing the method on a representative series of toy models, for which the model to reconstruct is already known. In contrast, it is not obvious how to prove that a given tool for content creation brings a new contribution. Our strategy to tackle the problem is threefold:

- Most of our contributions will address the design of new models and algorithms for geometry and animation. Validating them will be done, as usual in Computer Graphics, by showing for instance that our method solves a problem never solved before, that the model is more general, or the computations more efficient, than using previous methods.
- Interaction for interactive content creation & editing will rely as much as possible on preliminary user studies telling us about user expectations, and on interaction paradigms and design principles already identified and validated by the HCI community. When necessary, we intend to develop as well new interaction paradigms and devices (such as the hand-navigator we are currently experimenting) and validate them through user studies. All this interaction design work will be done in collaboration with the HCI community. We already set up a long term partnership with the IIHM group from LIG in Grenoble, through the INTUACTIVE project at Grenoble INP (2011-2014) which involves co-advised students, and through the co-direction of the action "Authoring Augmented Reality" of the larger Labex PERSYVAL project (2012 – 2020).
- Lastly, working on specific applications in the domains we listed in Section 3 is essential for validation since it will give us some test beds for real-size applications. The expert users involved will be able to validate the use of our new design framework compared to their usual pipeline, both in terms of increased efficiency, and of satisfaction with new functionalities and final result. In addition to our work with scientific and industrial partners, we are establishing collaborations with the Ecole Nationale Supérieure des Arts Décoratifs (ENSAD Paris, Prof Pierre Hénon) and with the Ecole Nationale Supérieure Louis Lumière (Prof. Pascal Martin) for the evaluation of our ongoing work in shape and motion design, and on virtual cinematography.

IMARA Project-Team

3. Scientific Foundations

3.1. Vehicle guidance and autonomous navigation

Participants: Fawzi Nashashibi, Evangeline Pollard, Benjamin Lefaudeux, Hao Li, Paulo Lopes Resende, Guillaume Tréhard, Pierre Merdrignac, Zayed Alsayed.

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

Either for driver assistance or for fully automated guided vehicles 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 and cooperative perception.

3.1.2. 3D environment mapping

Participants: Fawzi Nashashibi, Hao Li, Benjamin Lefaudeux, Paulo Lopes Resende.

In the past few years, we've 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.

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 other approach, originality relies in 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 also worked in the past on an original approach for 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, which corresponds to the intersection of the constraint sets. The construction of convex property sets is based on the various properties of the field to be estimated. In most stereo vision applications, the disparity map should be smooth in homogeneous areas while keeping sharp edges. This can be achieved with the help of a suitable regularization constraint. We propose to use the Total Variation information as a regularization constraint, which avoids oscillations while preserving field discontinuities around object edges.

The algorithm we developed to solve the estimation disparity problem has a block-iterative structure. This allows a wide range of constraints to be easily incorporated, possibly taking advantage of parallel computing architectures. This efficient algorithm allowed us to combine the Total Variation constraint with additional convex constraints so as to smooth homogeneous regions while preserving discontinuities.

Presently, we are currently working on an original stereo-vision based SLAM technique, 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 [1], and is built on linear algebra (through Inria's Eigen library), RANSAC 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: Fawzi Nashashibi, Hao Li, Evangeline Pollard, Benjamin Lefaudeux, Pierre Merdrignac.

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. IMARA 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

[1] http://www.youtube.com/watch?v=obH9Z2uOMBI

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's developments permitted the development of the basic hardware pieces that ensures the well functioning of the embedded architecture including perception sensors, communication devices and processing tools. The final architecture was relying on the *SensorHub* presented in year 2010 report and demonstrated several times in year 2011 (ITS World Congress, workshop "The automation for urban transport" in La Rochelle...)

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.

Both methods should be experimentally tested on IMARA veicles in 2013.

3.1.4. Planning and executing vehicle actions

Participants: Plamen Petrov, Joshué Pérez Rastelli, Fawzi Nashashibi, Philippe Morignot, Paulo Lopes Resende, Mohamed Marouf.

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've been focusing on the generation of geometric trajectories as a result of a manoeuvre selection process using grid-based rating technique or fuzzy technique. For high speed vehicles, Partial Motion Planning techniques we tested revealed their limitation 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 DLR's JointSystem demonstrator used in the European project HAVEit as well as in IMARA's electrical vehicle prototype used in the French project ABV. HAVEit was also the opportunity for IMARA 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 IMARA's cybercars. However, for the low speed cybercars that have pre-defined itineraries and basic manoeuvres it was necessary to develop a more adapted planning and control system. Therefore, we've 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, controlling the vehicles include also advanced platooning, automated parking, automated docking, etc. Foreach functionality a dedicated control algorithm was designed (see publication of previous years). Today, IMARA is also investigating the opportunity of fuzzy-based control for specific manoeuvres. First results have been recently obtained for reference trajectory following in roundabouts and normal straight roads.

3.2. V2V and V2I Communications for ITS

Participants: Thierry Ernst, Oyunchimeg Shagdar, Gérard Le Lann, Manabu Tsukada, Thouraya Toukabri, Satoru Noguchi, Ines Ben Jemaa, Mohammad Abu Alhoul, Fawzi Nashashibi, Arnaud de la Fortelle.

Wireless communications is 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, IMARA has been working on medium access control design and visible light communications especially for highly dense scenarios. The works have been carried out considering vehicles' behavior such as vehicles' merging and 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. In the context of IPv6, we have been tackling research issues of combinations of MANET and NEMO and Multihoming in Nested Mobile Networks with Route Optimization.

The wireless channel and topology dynamics are the characteristics that require great research challenge in 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. Multihoming in nested mobile networks with route optimization

Participants: Manabu Tsukada, Thierry Ernst.

Network mobility has the particularity of allowing recursive mobility, i.e. where a mobile node is attached to another mobile node (e.g. a PDA is attached to the in-vehicle IP network). This is referred to as nested mobility and brings a number of research issues in terms of routing efficiency. Another issue under such mobility configurations is the availability of multiple paths to the Internet (still in the same example, the PDA has a 3G interface and the in-vehicle network has some dedicated access to the Internet) and its appropriate selection.

3.2.2. Service discovery

Participants: Satoru Noguchi, Thierry Ernst.

Vehicles in a close vicinity need to discover what information can be made available to other vehicles (e.g. road traffic conditions, safety notification for collision avoidance). We are investigating both push and pull approaches and the ability of these mechanisms to scale to a large number of vehicles and services on offer.

3.2.3. Geographic multicast addressing and routing

Participants: Ines Ben Jemaa, Oyunchimeg Shagdar, Thierry Ernst, Arnaud de La Fortelle, Fawzi Nashashibi.

Many ITS applications such as fleet management require multicast data delivery. Existing works on this subject tackle 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) to ensure vehicular multicast group reachability through the infrastructure network, defines a distributed and efficient geographic multicast auto-addressing mechanism and ii) to allow simple and efficient data delivery introduces a simplified approach that locally manages the group membership and distributes the packets among them.

3.2.4. 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 that VLC technology has a short history in comparing with other communication technologies, the infrastructure availability and the presence of the congestion in wireless communication channels are proposing VLC technology as reliable and supportive technology which can takeoff some loads of the wireless radio communication. First objective of this work is develop analytical model of VLC to understand its characteristics and limitation. The second objective of this work is to design vehicle platooning control using VLC. In platooning control, a corporation between control and communication is strongly required in order guarantee the platoon's stability (e.g. string stability problem). For this purpose we work on VLC model platooning scenario, to permit 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 to the main Simulink platooning model will be responsible for deciding the availability of the Line-of-Sight for different trajectory's curvatures, which mean the capability of using light communication between each two vehicles in the platooning queue, at the same time the model will calculate all the required parameters acquired from each vehicle controller.

3.2.5. Access point selection

Participant: Oyunchimeg Shagdar.

While 5.9 GHz radio frequency band is dedicated to ITS applications, there is not much known how the channel and network behave in mobile scenarios. 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 access point selection method to achieve high speed V2I communications.

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 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 omnidirectional 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 omnidirectional (360°) communications and unidirectional communications, implemented out of very-short range antennas of very small beamwidth. 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 [7] and [31].

3.3.2. Cohorts, N2N communications, and safety in the presence of telemetry failures

In [7] and [31], 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 increase 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 [49], 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. Managing the system (via probabilistic modeling)

Participants: Guy Fayolle, Cyril Furtlehner, Yufei Han, Arnaud de La Fortelle, Jean-Marc Lasgouttes, Victorin Martin.

The research on the management of the transportation system is a natural continuation of the research of the Preval team, which joined IMARA in 2007. For many years, the members of this team (and of its ancestor Meval) have been working on understanding random systems of various origins, mainly through the definition and solution of mathematical models. The traffic modeling field is very fertile in difficult problems, and it has been part of the activities of the members of Preval since the times of the Praxitèle project.

Following this tradition, the roadmap of the group is to pursue basic research on probabilistic modeling with a clear slant on applications related to LaRA activities. A particular effort is made to publicize our results among the traffic analysis community, and to implement our algorithms whenever it makes sense to use them in traffic management. Of course, as aforementioned, these activities in no way preclude the continuation of the methodological work achieved in the group for many years in various fields: random walks in Z_+^n ([1], [2], [5]), large deviations, birth and death processes on trees, particle systems. The reader is therefore encouraged to read the recent activity reports for the Preval team for more details.

In practice, the group explores the links between large random systems and statistical physics, since this approach proves very powerful, both for macroscopic (fleet management [4]) and microscopic (car-level description of traffic, formation of jams) analysis. The general setting is mathematical modeling of large systems (mostly stochastic), without any a priori restriction: networks [3], random graphs or even objects coming from biology. When the size or the volume of those structures grows (this corresponds to the so-called thermodynamical limit), one aims at establishing a classification based on criteria of a twofold nature: quantitative (performance, throughput, etc) and qualitative (stability, asymptotic behavior, phase transition, complexity).

3.4.1. Exclusion processes

One of the simplest basic (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.

Most of these generalizations lead to models that are obviously difficult to solve and require upstream theoretical studies. Some of them models have already been investigated by members of the group, and they are part of wide ongoing research.

3.4.2. Message passing algorithms

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. Using the Ising model, together with the Belief Propagation algorithm very popular in the computer science community, 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 properties of the Bethe approximation of Ising models:

- determine the effect of the various variants of BP (in terms of normalization or changes to the Bethe free energy) on the fixed points and their stability;

- find the best way to inject real-valued data in an Ising model with binary variables;
- build macroscopic variables that measure the overall state of the underlying graph, in order to improve the local propagation of information;
- make the underlying model as sparse as possible, in order to improve BP convergence and quality.

IMEDIA2 Team

3. Scientific Foundations

3.1. Introduction

We group the existing problems in the domain of content-based image indexing and retrieval in the following themes: image indexing and efficient search in image collections, pattern recognition and personalization. In the following we give a short introduction to each of these themes.

3.2. Modeling, construction and structuring of the feature space

Participants: Vera Bakic, Nozha Boujemaa, Esma Elghoul, Hervé Goëau, Amel Hamzaoui, Sofiene Mouine, Olfa Mzoughi, Saloua Ouertani-Litayem, Mohamed Riadh Trad, Anne Verroust-Blondet, Itheri Yahiaoui, Zahraa Yasseen.

The goal of IMEDIA2 team is to provide the user with the ability to do content-based search into image databases in a way that is both intelligent and intuitive to the users. When formulated in concrete terms, this problem gives birth to several mathematical and algorithmic challenges.

To represent the content of an image, we are looking for a representation that is both compact (less data and more semantics), relevant (with respect to the visual content and the users) and fast to compute and compare. The choice of the feature space consists in selecting the significant *features*, the *descriptors* for those features and eventually the encoding of those descriptors as image *signatures*.

We deal both with generic databases, in which images are heterogeneous (for instance, search of Internet images), and with specific databases, dedicated to a specific application field. The specific databases are usually provided with a ground-truth and have an homogeneous content (leaf images, for example)

We must not only distinguish generic and specific signatures, but also local and global ones. They correspond respectively to queries concerning parts of pictures or entire pictures. In this case, we can again distinguish approximate and precise queries. In the latter case one has to be provided with various descriptions of parts of images, as well as with means to specify them as regions of interest. In particular, we have to define both global and local similarity measures.

When the computation of signatures is over, the image database is finally encoded as a set of points in a high-dimensional space: the feature space.
A second step in the construction of the index can be valuable when dealing with very high-dimensional feature spaces. It consists in pre-structuring the set of signatures and storing it efficiently, in order to reduce access time for future queries (trade-off between the access time and the cost of storage). In this second step, we have to address problems that have been dealt with for some time in the database community, but arise here in a new context: image databases. Today's scalability issues already put brake on growth of multi-media search engines. The space created by the massive amounts of existing multimedia files greatly exceeds the area searched by today's major engines. Consistent breakthroughs are therefore urgent if we don't want to be lost in data space in ten years. We believe that reducing algorithm complexity remains the main key. Whatever the efficiency of the implementation or the use of powerful hardware and distributed architectures, the ability of an algorithm to scale-up is strongly related to its time and space complexities. Nowadays, efficient multimedia search engines rely on various high level tasks such as content-based search, navigation, knowledge discovery, personalization, collaborative filtering or social tagging. They involve complex algorithms such as similarity search, clustering or machine learning, on heterogeneous data, and with heterogeneous metrics. Some of them still have quadratic and even cubic complexities so that their use in the large scale is not affordable if no fundamental research is performed to reduce their complexities. In this way, efficient and generic high-dimensional similarity search structures are essential for building scalable content-based search systems. Efficient search requires a specific structuring of the feature space (multidimensional indexing, where indexing is understood as data structure) for accelerating the access to collections that are too large for the central memory.

3.3. Pattern recognition and statistical learning

Participants: Nozha Boujemaa, Michel Crucianu, Donald Geman, Wajih Ouertani, Asma Rejeb Sfar.

Statistical learning and classification methods are of central interest for content-based image retrieval. We consider here both supervised and unsupervised methods. Depending on our knowledge of the contents of a database, we may or may not be provided with a set of *labeled training examples*. For the detection of *known* objects, methods based on hierarchies of classifiers have been investigated. In this context, face detection was a main topic, as it can automatically provide a high-level semantic information about video streams. For a collection of pictures whose content is unknown, e.g. in a navigation scenario, we are investigating techniques that adaptively identify homogeneous clusters of images, which represent a challenging problem due to feature space configuration.

Object detection is the most straightforward solution to the challenge of content-based image indexing. Classical approaches (artificial neural networks, support vector machines, etc.) are based on induction, they construct generalization rules from training examples. The generalization error of these techniques can be controlled, given the complexity of the models considered and the size of the training set.

Our research on object detection addresses the design of invariant kernels and algorithmically efficient solutions as well as boosting method for similarity learning. We have developed several algorithms for face detection based on a hierarchical combination of simple two-class classifiers. Such architectures concentrate the computation on ambiguous parts of the scene and achieve error rates as good as those of far more expensive techniques.

Unsupervised clustering techniques automatically define categories and are for us a matter of visual knowledge discovery. We need them in order to:

- Solve the "page zero" problem by generating a visual summary of a database that takes into account all the available signatures together.
- Perform image segmentation by clustering local image descriptors.
- Structure and sort out the signature space for either global or local signatures, allowing a hierarchical search that is necessarily more efficient as it only requires to "scan" the representatives of the resulting clusters.

Given the complexity of the feature spaces we are considering, this is a very difficult task. Noise and class overlap challenge the estimation of the parameters for each cluster. The main aspects that define the clustering process and inevitably influence the quality of the result are the clustering criterion, the similarity measure and the data model.

IN-SITU Project-Team

3. Scientific Foundations

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 [40].

LAGADIC Project-Team

3. Scientific Foundations

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:

$$\mathbf{s} = \mathbf{s}(\mathbf{x}(\mathbf{p}(t)), \mathbf{a}) \tag{111}$$

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:

$$\dot{\mathbf{s}} = \frac{\partial \mathbf{s}}{\partial \mathbf{p}} \dot{\mathbf{p}} = \mathbf{L}_{\mathbf{s}} \mathbf{v} \tag{112}$$

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:

$$\mathbf{v}_c = -\lambda \widehat{\mathbf{L}_{\mathbf{s}}}^{+} (\mathbf{s} - \mathbf{s}^*) - \widehat{\mathbf{L}_{\mathbf{s}}}^{+} \widehat{\frac{\partial \mathbf{s}}{\partial t}} \tag{113}$$

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. Scientific Foundations

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.

MAGRIT Project-Team

3. Scientific Foundations

3.1. Camera calibration and registration

One of the most basic problems currently limiting Augmented Reality 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, and the user should walk anywhere he pleases.

For several years, the Magrit project has been aiming at developing on-line and marker-less methods for camera pose computation. We have especially proposed a real-time system for camera tracking designed for indoor scenes [1]. The main difficulty with on-line tracking is to ensure robustness of the process. 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 developed a method which combines the advantage of move-matching methods and model-based methods by using a piecewise-planar model of the environment. This methodology can be used in a wide variety of environments: indoor scenes, urban scenes .... We are also concerned with the development of methods for camera stabilization. Indeed, statistical fluctuations in the viewpoint computations lead to unpleasant jittering or sliding effects, especially when the camera motion is small. We have proved that the use of model selection allows us to noticeably improve the visual impression and to reduce drift over time.

The success of pose computation largely depends on the quality of the matching stage over the sequence. Research are conducted in the team on the use of probabilistic methods to establish robust correspondences of features over time. The use of *a contrario* decision is under investigation to achieve this aim [3]. We especially address the complex case of matching in scenes with repeated patterns which are common in urban scenes. We also consider learning based techniques to improve the robustness of the matching stage.

Another way 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. Each technology approach has limitations: on the one hand, rapid head motions cause image features to undergo large motion between frames that may cause visual tracking to fail. On the other hand, inertial sensors response is largely independent from the user's motion but their accuracy is bad and their response is sensitive to metallic objects in the scene. In past works [1], we have proposed a system that makes an inertial sensor cooperate with the camera-based system in order to improve the robustness of the AR system to abrupt motions of the users, especially head motions. This work contributes to the reduction of the constraints on the users and the need to carefully control the environment during an AR application. Ongoing research on such hybrid systems are under consideration in our team with the aim to improve the accuracy of reconstruction techniques as well as to obtain dynamic models of organs in medical applications.

Finally, it must be noted that the registration problem must be addressed from the specific point of view of augmented reality: the success and the acceptance of an AR application does not only depend on the accuracy of the pose computation but also on the visual impression of the augmented scene. The search for the best compromise between accuracy and perception is therefore an important issue in this project. This research topic has been addressed in our project both in classical AR and in medical imaging in order to choose the camera model, including intrinsic parameters, which describes at best the considered camera.

3.2. Scene 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 occlusion and to compute light reflexions between the real and the virtual objects. 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 according to the application. Within the team we have developed interactive in-situ modeling techniques dedicated to classical AR applications. We also developed off-line multimodal techniques dedicated to AR medical applications.

In-situ modeling
Most automatic techniques aim at reconstructing a sparse and thus unstructured set of points of the scene. Such models are obviously not appropriate to perform interaction with the scene. In addition, they are incomplete in the sense that they may omit features which are important for the accuracy of the pose recovered from 2D/3D correspondences. We have thus investigated interactive techniques with the aim of obtaining reliable and structured models of the scene. The goal of our approach is to develop immersive and intuitive interaction techniques which allow for scene modeling during the application [7].

Multimodal modeling 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 are 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 and the 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 15 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. An accurate definition of the target is a parameter of great importance for the success of the treatment. 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 Shacra EPI, 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 [4]. Multi-modality techniques for reconstruction are also considered within the european ASPI project, the aim of which is to build a dynamic model of the vocal tract from various images modalities (MRI, ultrasound, video) and magnetic sensors.

MAIA Project-Team

3. Scientific Foundations

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 research trends:

(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. We remark that they are not exhaustive. In parts of our work, we also consider the difference between discrete/continuous problems, and centralized/decentralized problems. The complexity of solving the problem – both in theory and in practice – depends crucially on where the problem resides in this categorization. In many applications, not one but several points in the categorization make sense: simplified versions of the problem can be solved much more effectively and thus serve for the generation of *some* – if possibly sub-optimal – action strategy in a more feasible manner. Of course, the application as such may also come in different facets.

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** [51]. 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 allowing to model complex problems 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**) [52] 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 [57]); (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-POMDPs, 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.

3.1.3. Project-team positioning

Within Inria, the most closely related teams are TAO and Sequel. TAO works on evolutionary computation (EC) and statistical machine learning (ML), and their combination. Sequel works on ML, with a theoretical focus combining CS and applied maths. The main difference is that TAO and Sequel consider particular algorithmic frameworks that can, amongst others, be applied to Planning and Reinforcement Learning, whereas we revolve around Planning and Reinforcement Learning as the core problems to be tackled, with whichever framework suitable.

In France, we have recently begun collaborating with the IMS Team of Supélec Metz, notably with O. Pietquin and M. Geist who have a great expertise in approximate techniques for MDPs. We have have links with the MAD team of the BIA unit of the INRA at Toulouse, led by R. Sabbadin. They also use MDP related models and are interested in solving large size problems, but they are more driven by applications (mostly agricultural) than we are. In Paris, the Animat Lab, that was a part of the LIP6 and is now attached to the ISIR, has done some interesting works on factored Markov Decision Problems and POMDPs. Like us, their main goal was to tackle problems with large state space.

In Europe, the IDSIA Lab at Lugano (Switzerland) has brought some interesting ideas to the field of MDP (meta-learning, subgoal discovery) but seems now more interested in a *Universal Learner*. In Osnabrück (Germany), the Neuroinformatic group works on efficient reinforcement learning with a specific interest in the application to robotics. For deterministic planning, the most closely related groups are located in Freiburg (Germany), Glasgow (UK), and Barcelona (Spain). We have active collaborations with all of these.

In the rest of the world, the most important groups regarding MDPs can be found at Brown University, Rutgers Univ. (M. Littman), Univ. of Toronto (C. Boutilier), MIT AI Lab (L. Kaelbling, D. Bertsekas, J.Tsitsiklis), Stanford Univ., CMU, Univ. of Alberta (R. Sutton), Univ. of Massachusetts at Amherst (S. Zilberstein, A. Barto), *etc*. A major part of their work is aimed at making Markov Decision Process based tools work on real life problems and, as such, our scientific concerns meet theirs. For deterministic planning, important related groups and collaborators are to be found at NICTA (Canberra, Australia) and at Cornell University (USA).

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 more clearly 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

To represent these complex systems, we made the choice to use 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 model 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 of 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

Currently, there is no agreement on a formal definition of a multi-agent system. 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 mechanism 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.

3.2.4. Project-team positioning

Building a reactive multi-agent system consists in defining a set (generally a large number) of simple and reactive agents within a shared environment (physical or virtual) in which they move, act and interact with each other. Our interest in these systems is that, in spite of their simple definition at the agent level, they produce coherent and coordinated behavior at a global scale. The properties that they may exhibit, such as robustness and adaptivity explain why their study has been growing in the last decade (in the broader context of "complex systems").

Our work on such problems is characterized by five research trends: (A) *Defining a formal framework for describing and studying these systems*, (B) *Developing and understanding reactive multi-agent systems*, (C) *Analysing and proving properties*, (D) *Deploying these systems on typical distributed architectures such as swarms of robots, FPGAs, GPUs and sensor networks*, (E) *Transferring our results in applications.*

Multi-agent System is an active area of research in Artificial Intelligence and Complex Systems. Our research fits well into the international research context, and we have made and are making a variety of significant contributions both in theoretical and practical issues. Concerning multi-agent simulation and formalization, we compete or collaborate in France with S. Hassas in LIESP (Lyon), CERV (Brest), IREMIA (la Réunion), Ibisc (Evry), Lirmm (Montpellier), Irit (Toulouse), A. Drogoul (IRD, Bondy) and abroad with F. Zambonelli (Univ. Modena, Italy) A. Deutsch (Dresden, Germany), D. Van Parunak (Vector research, USA), P. Valkenaers, D. Weyns (Univ. Leuven, Belgium), etc. Regarding our work on swarm robotics we have common objectives with the DISAL [3] EPFL Laboratory, the Bristol Robotics Laboratory, the Distributed Robotics Laboratory at MIT, the team of W. & D. Spears at Wyoming university, the Pheromone Robotics project at HRL Lab. [4], the FlockBots project at GMU[5], the team of G. Théraulaz at CNRS-Toulouse and the teams of J.-L. Deneubourg and M. Dorigo at ULB (Bruxelles).

[3] Distributed Intelligent Systems and Algorithms Laboratory including EPFL Swarm-Intelligent Systems Group (SWIS) founded in 2003 and the Collective Robotics Group (CORO) founded in 2000 at California Institute of Technology USA

[4] HRL, Information and systems sciences Lab (ISSL), Malibu CA, USA (D. Payton)

[5] George Mason University, Eclab, USA (L. Panait, S. Luke)

MANAO Team

3. Scientific Foundations

3.1. Related Scientific Domains

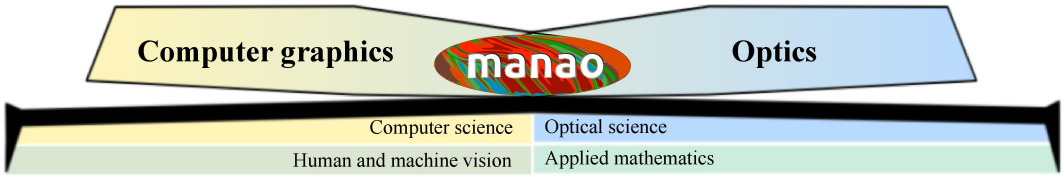


Figure 4. Related scientific domains of the MANAO project.

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 4) 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 [43] 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 [55], [56] and display [54] 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 6.3). 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 [45] or differential analysis [74], 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) 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 Figure3 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 [5], 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 [51] in order to develop tailored strategies [88]. For instance, starting from simple direct lighting, more complex phenomena

have been progressively introduced: first diffuse indirect illumination [49], [81], then more generic inter-reflections [58], [43] and volumetric scattering [78], [40]. 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 [39]. 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 [80] but are difficult to extend to non-piecewise-constant data [83]. More recently, researches prefer the use of Spherical Radial Basis Functions [86] or Spherical Harmonics [73]. For more complex data, such as reflective properties (e.g., BRDF [68], [59] - 4D), ray-space (e.g., Light-Field [65] - 4D), spatially varying reflective properties (6D - [77]), new models, and representations are still investigated such as rational functions [19] or dedicated models [28] and parameterizations [79], [84]. 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 [19].

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 [74] and frequency analysis [45]). Such an approach has leaded us to exhibit the influence of surfaces flows (depth and normal gradients) into lighting pattern deformation (see Figure 8). For a **human observer**, this correspond to one recent trend in computer graphics that takes into account the human visual systems [47] 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.

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 [65], [54]. We consider projecting systems and surfaces [35], for personal use, virtual reality and augmented reality [31]. 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 [2], [48]. These resulting systems

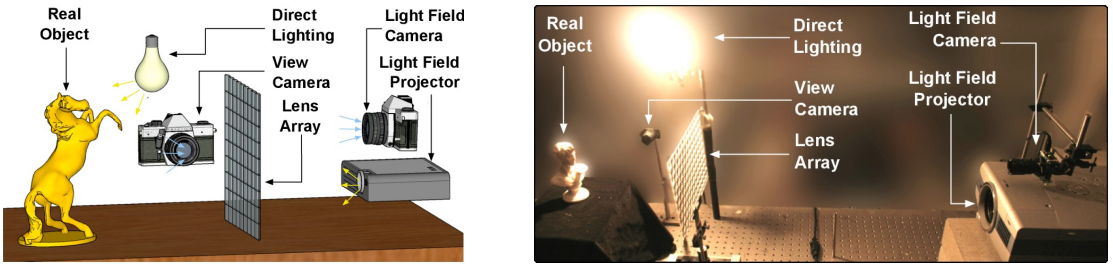


Figure 5. Light-Field transfer: global illumination between real and synthetic objects [38]

have to acquire the different phenomena described in Axis 1 and to display them, in an efficient manner [52], [29], [53], [56]. 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 [57].

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 [56], [72]). Similarly, increasing the number of viewpoints is a way toward the creation of full 3D displays [65]. 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., [71], [89]).

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. [38]. 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 [63] are one of the key technologies to develop such acquisition (e.g., Light-Field camera [1] [57] and acquisition of light-sources [2]), projection systems (e.g., auto-stereoscopic displays). Such an approach is versatile and may be applied to improve classical optical instruments [62]. More generally, by designing unified optical and digital systems [69], 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

[1] Lytro, http://www.lytro.com/

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 [19]. Second, we have to integrate signals from multiple sensors (such as GPS, accelerometer, ...) to prevent some computation (e.g., [17]). Finally, the experience of the group in surface modeling help the design of optical surfaces [60] 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 [70] (b) Shadows [30] (c) Shape enhancement [8] (d) Shape depiction [27]

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.

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 [9], [61] or stylized shading [33],[8] 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 [15] 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 [10], motion blur [46], depth of field [82], 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 [3], [44]. 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 [37]. Independent image processing for each viewpoint may disturb the feeling of depth by introducing inconsistent information in each images. Finally, more dedicated displays [54] 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.

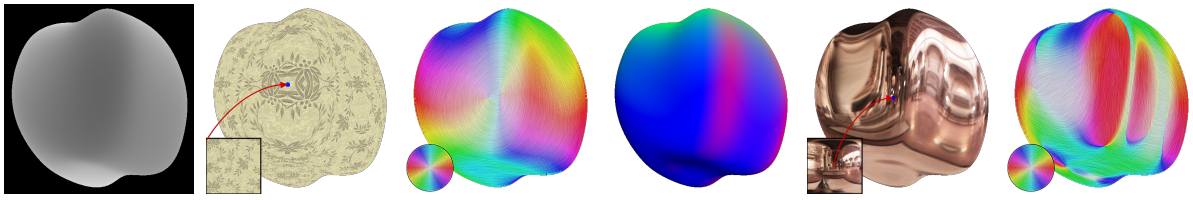
Results: High-level primitives and hybrid representations for appearance and shape.

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 [4], [1]. 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., [6]): 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 [6]), 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 [21] (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 Team

3. Scientific Foundations

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.

METISS Project-Team

3. Scientific Foundations

3.1. Introduction

probabilistic modeling, statistical estimation, bayesian decision theory gaussian mixture modeling, Hidden Markov Model, adaptive representation, redundant system, sparse decomposition, sparsity criterion, source separation

Probabilistic approaches offer a general theoretical framework [92] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [89], the probabilistic framework indeed provides a solid formalism which makes it possible to formulate various problems of segmentation, detection and classification. Coupled to statistical approaches, the probabilistic paradigm makes it possible to easily adapt relatively generic tools to various applicative contexts, thanks to estimation techniques for training from examples.

A particularily productive family of probabilistic models is the Hidden Markov Model, either in its general form or under some degenerated variants. The stochastic framework makes it possible to rely on well-known algorithms for the estimation of the model parameters (EM algorithms, ML criteria, MAP techniques, ...) and for the search of the best model in the sense of the exact or approximate maximum likelihood (Viterbi decoding or beam search, for example).

More recently, Bayesian networks [94] have emerged as offering a powerful framework for the modeling of musical signals (for instance, [90], [95]).

In practice, however, the use of probabilistic models must be accompanied by a number of adjustments to take into account problems occurring in real contexts of use, such as model inaccuracy, the insufficiency (or even the absence) of training data, their poor statistical coverage, etc...

Another focus of the activities of the METISS research group is dedicated to sparse representations of signals in redundant systems [93]. The use of criteria of sparsity or entropy (in place of the criterion of least squares) to force the unicity of the solution of a underdetermined system of equations makes it possible to seek an economical representation (exact or approximate) of a signal in a redundant system, which is better able to account for the diversity of structures within an audio signal.

The topic of sparse representations opens a vast field of scientific investigation : sparse decomposition, sparsity criteria, pursuit algorithms, construction of efficient redundant dictionaries, links with the non-linear approximation theory, probabilistic extensions, etc... and more recently, compressive sensing [88]. The potential applicative outcomes are numerous.

This section briefly exposes these various theoretical elements, which constitute the fundamentals of our activities.

3.2. Probabilistic approach

probability density function, gaussian model, gaussian mixture model, Hidden Markov Model, Bayesian network, maximum likelihood, maximum a posteriori, EM algorithm, inference, Viterbi algorithm, beam search, classification, hypotheses testing, acoustic parameterisation

For several decades, the probabilistic approaches have been used successfully for various tasks in pattern recognition, and more particularly in speech recognition, whether it is for the recognition of isolated words, for the retranscription of continuous speech, for speaker recognition tasks or for language identification. Probabilistic models indeed make it possible to effectively account for various factors of variability occuring in the signal, while easily lending themselves to the definition of metrics between an observation and the model of a sound class (phoneme, word, speaker, etc...).

3.2.1. Probabilistic formalism and modeling

The probabilistic approach for the representation of an (audio) class X relies on the assumption that this class can be described by a probability density function (PDF) $P(.|X)$ which associates a probability $P(Y|X)$ to any observation Y.

In the field of speech processing, the class X can represent a phoneme, a sequence of phonemes, a word from a vocabulary, or a particular speaker, a type of speaker, a language, ... Class X can also correspond to other types of sound objects, for example a family of sounds (word, music, applause), a sound event (a particular noise, a jingle), a sound segment with stationary statistics (on both sides of a rupture), etc.

In the case of audio signals, the observations Y are of an acoustical nature, for example vectors resulting from the analysis of the short-term spectrum of the signal (filter-bank coefficients, cepstrum coefficients, time-frequency principal components, etc.) or any other representation accounting for the information that is required for an efficient separation of the various audio classes considered.

In practice, the PDF P is not accessible to measurement. It is therefore necessary to resort to an approximation $\widehat{P}$ of this function, which is usually refered to as the likelihood function. This function can be expressed in the form of a parametric model.

The models most used in the field of speech and audio processing are the Gaussian Model (GM), the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM). But recently, more general models have been considered and formalised as graphical models.

Choosing a particular family of models is based on a set of considerations ranging from the general structure of the data, some knowledge on the audio class making it possible to size the model, the speed of calculation of the likelihood function, the number of degrees of freedom of the model compared to the volume of training data available, etc.

3.2.2. Statistical estimation

The determination of the model parameters for a given class is generally based on a step of statistical estimation consisting in determining the optimal value for model parameters.

The Maximum Likelihood (ML) criterion is generally satisfactory when the number of parameters to be estimated is small w.r.t. the number of training observations. However, in many applicative contexts, other estimation criteria are necessary to guarantee more robustness of the learning process with small quantities of training data. Let us mention in particular the Maximum a Posteriori (MAP) criterion which relies on a prior probability of the model parameters expressing possible knowledge on the estimated parameter distribution for the class considered. Discriminative training is another alternative to these two criteria, definitely more complex to implement than the ML and MAP criteria.

In addition to the fact that the ML criterion is only one particular case of the MAP criterion, the MAP criterion happens to be experimentally better adapted to small volumes of training data and offers better generalization capabilities of the estimated models (this is measured for example by the improvement of the classification performance and recognition on new data). Moreover, the same scheme can be used in the framework of incremental adaptation, i.e. for the refinement of the parameters of a model using new data observed for instance, in the course of use of the recognition system.

3.2.3. Likelihood computation and state sequence decoding

During the recognition phase, it is necessary to evaluate the likelihood function of the observations for one or several models. When the complexity of the model is high, it is generally necessary to implement fast calculation algorithms to approximate the likelihood function.

In the case of HMM models, the evaluation of the likelihood requires a decoding step to find the most probable sequence of hidden states. This is done by implementing the Viterbi algorithm, a traditional tool in the field of speech recognition. However, when the acoustic models are combined with a syntagmatic model, it is necessary to call for sub-optimal strategies, such as beam search.

3.2.4. Bayesian decision

When the task to solve is the classification of an observation into one class among several closed-set possibilities, the decision usually relies on the maximum a posteriori rule.

In other contexts (for instance, in speaker verification, word-spotting or sound class detection), the problem of classification can be formulated as a binary hypotheses testing problem, consisting in deciding whether the tested observation is more likely to be pertaining to the class under test or not pertaining to it. In this case, the decision consists in acceptance or rejection, and the problem can be theoretically solved within the framework of Bayesian decision by calculating the ratio of the PDFs for the class and the non-class distributions, and comparing this ratio to a decision threshold.

In theory, the optimal threshold does not depend on the class distribution, but in practice the quantities provided by the probabilistic models are not the true PDFs, but only likelihood functions which approximate the true PDFs more or less accurately, depending on the quality of the model of the class.

The optimal threshold must be adjusted for each class by modeling the behaviour of the test on external (development) data.

3.2.5. Graphical models

In the past years, increasing interest has focused on graphical models for multi-source audio signals, such as polyphonic music signals. These models are particularly interesting, since they enable a formulation of music modelisation in a probabilistic framework.

It makes it possible to account for more or less elaborate relationship and dependencies between variables representing multiple levels of description of a music piece, together with the exploitation of various priors on the model parameters.

Following a well-established metaphore, one can say that the graphical model expresses the notion of modularity of a complex system, while probability theory provides the glue whereby the parts are combined. Such a data structure lends itself naturally to the design of efficient general-purpose algorithms.

The graphical model framework provides a way to view a number of existing models (including HMMs) as instances of a common formalism and all of them can be addressed via common machine learning tools.

A first issue when using graphical models is the one of the model design, i.e. the chosen variables for parameterizing the signal, their priors and their conditional dependency structure.

The second problem, called the inference problem, consists in estimating the activity states of the model for a given signal in the maximum a posteriori sense. A number of techniques are available to achieve this goal (sampling methods, variational methods belief propagation, ...), whose challenge is to achieve a good compromise between tractability and accuracy [94].

3.3. Sparse representations

wavelet, dictionary, adaptive decomposition, optimisation, parcimony, non-linear approximation, pursuit, greedy algorithm, computational complexity, Gabor atom, data-driven learning, principal component analysis, independant component analysis

Over the past decade, there has been an intense and interdisciplinary research activity in the investigation of sparsity and methods for sparse representations, involving researchers in signal processing, applied mathematics and theoretical computer science. This has led to the establishment of sparse representations as a key methodology for addressing engineering problems in all areas of signal and image processing, from the data acquisition to its processing, storage, transmission and interpretation, well beyond its original applications in enhancement and compression. Among the existing sparse approximation algorithms, L1-optimisation principles (Basis Pursuit, LASSO) and greedy algorithms (e.g., Matching Pursuit and its variants) have in particular been extensively studied and proved to have good decomposition performance, provided that the sparse signal model is satisfied with sufficient accuracy.

The large family of audio signals includes a wide variety of temporal and frequential structures, objects of variable durations, ranging from almost stationary regimes (for instance, the note of a violin) to short transients (like in a percussion). The spectral structure can be mainly harmonic (vowels) or noise-like (fricative consonants). More generally, the diversity of timbers results in a large variety of fine structures for the signal and its spectrum, as well as for its temporal and frequential envelope. In addition, a majority of audio signals are composite, i.e. they result from the mixture of several sources (voice and music, mixing of several tracks, useful signal and background noise). Audio signals may have undergone various types of distortion, recording conditions, media degradation, coding and transmission errors, etc.

Sparse representations provide a framework which has shown increasingly fruitful for capturing, analysing, decomposing and separating audio signals

3.3.1. Redundant systems and adaptive representations

Traditional methods for signal decomposition are generally based on the description of the signal in a given basis (i.e. a free, generative and constant representation system for the whole signal). On such a basis, the representation of the signal is unique (for example, a Fourier basis, Dirac basis, orthogonal wavelets, ...). On the contrary, an adaptive representation in a redundant system consists of finding an optimal decomposition of the signal (in the sense of a criterion to be defined) in a generating system (or dictionary) including a number of elements (much) higher than the dimension of the signal.

Let y be a monodimensional signal of length T and D a redundant dictionary composed of $N > T$ vectors g_i of dimension T.

$$y = [y(t)]_{1 \leq t \leq T} \qquad D = \{g_i\}_{1 \leq i \leq N} \quad \text{with} \quad g_i = [g_i(t)]_{1 \leq t \leq T}$$

If D is a generating system of R^T, there is an infinity of exact representations of y in the redundant system D, of the type:

$$y(t) = \sum_{1 \leq i \leq N} \alpha_i g_i(t)$$

We will denote as $\alpha = \{\alpha_i\}_{1 \leq i \leq N}$, the N coefficients of the decomposition.

The principles of the adaptive decomposition then consist in selecting, among all possible decompositions, the best one, i.e. the one which satisfies a given criterion (for example a sparsity criterion) for the signal under consideration, hence the concept of adaptive decomposition (or representation). In some cases, a maximum of T coefficients are non-zero in the optimal decomposition, and the subset of vectors of D thus selected are refered to as the basis adapted to y. This approach can be extended to approximate representations of the type:

$$y(t) = \sum_{1 \leq i \leq M} \alpha_{\phi(i)} g_{\phi(i)}(t) + e(t)$$

with $M < T$, where ϕ is an injective function of $[1, M]$ in $[1, N]$ and where $e(t)$ corresponds to the error of approximation to M terms of $y(t)$. In this case, the optimality criterion for the decomposition also integrates the error of approximation.

3.3.2. Sparsity criteria

Obtaining a single solution for the equation above requires the introduction of a constraint on the coefficients α_i. This constraint is generally expressed in the following form :

$$\alpha^* = \arg\min_{\alpha} F(\alpha)$$

Among the most commonly used functions, let us quote the various functions L_γ :

$$L_\gamma(\alpha) = \left[\sum_{1 \leq i \leq N} |\alpha_i|^\gamma \right]^{1/\gamma}$$

Let us recall that for $0 < \gamma < 1$, the function L_γ is a sum of concave functions of the coefficients α_i. Function L_0 corresponds to the number of non-zero coefficients in the decomposition.

The minimization of the quadratic norm L_2 of the coefficients α_i (which can be solved in an exact way by a linear equation) tends to spread the coefficients on the whole collection of vectors in the dictionary. On the other hand, the minimization of L_0 yields a maximally parsimonious adaptive representation, as the obtained solution comprises a minimum of non-zero terms. However the exact minimization of L_0 is an untractable NP-complete problem.

An intermediate approach consists in minimizing norm L_1, i.e. the sum of the absolute values of the coefficients of the decomposition. This can be achieved by techniques of linear programming and it can be shown that, under some (strong) assumptions the solution converges towards the same result as that corresponding to the minimization of L_0. In a majority of concrete cases, this solution has good properties of sparsity, without reaching however the level of performance of L_0.

Other criteria can be taken into account and, as long as the function F is a sum of concave functions of the coefficients α_i, the solution obtained has good properties of sparsity. In this respect, the entropy of the decomposition is a particularly interesting function, taking into account its links with the information theory.

Finally, let us note that the theory of non-linear approximation offers a framework in which links can be established between the sparsity of exact decompositions and the quality of approximate representations with M terms. This is still an open problem for unspecified redundant dictionaries.

3.3.3. Decomposition algorithms

Three families of approaches are conventionally used to obtain an (optimal or sub-optimal) decomposition of a signal in a redundant system.

The “Best Basis” approach consists in constructing the dictionary D as the union of B distinct bases and then to seek (exhaustively or not) among all these bases the one which yields the optimal decomposition (in the sense of the criterion selected). For dictionaries with tree structure (wavelet packets, local cosine), the complexity of the algorithm is quite lower than the number of bases B, but the result obtained is generally not the optimal result that would be obtained if the dictionary D was taken as a whole.

The “Basis Pursuit” approach minimizes the norm L_1 of the decomposition resorting to linear programming techniques. The approach is of larger complexity, but the solution obtained yields generally good properties of sparsity, without reaching however the optimal solution which would have been obtained by minimizing L_0.

The “Matching Pursuit” approach consists in optimizing incrementally the decomposition of the signal, by searching at each stage the element of the dictionary which has the best correlation with the signal to be decomposed, and then by subtracting from the signal the contribution of this element. This procedure is repeated on the residue thus obtained, until the number of (linearly independent) components is equal to the dimension of the signal. The coefficients α can then be reevaluated on the basis thus obtained. This greedy algorithm is sub-optimal but it has good properties for what concerns the decrease of the error and the flexibility of its implementation.

Intermediate approaches can also be considered, using hybrid algorithms which try to seek a compromise between computational complexity, quality of sparsity and simplicity of implementation.

3.3.4. Dictionary construction

The choice of the dictionary D has naturally a strong influence on the properties of the adaptive decomposition : if the dictionary contains only a few elements adapted to the structure of the signal, the results may not be very satisfactory nor exploitable.

The choice of the dictionary can rely on a priori considerations. For instance, some redundant systems may require less computation than others, to evaluate projections of the signal on the elements of the dictionary. For this reason, the Gabor atoms, wavelet packets and local cosines have interesting properties. Moreover, some general hint on the signal structure can contribute to the design of the dictionary elements : any knowledge on the distribution and the frequential variation of the energy of the signals, on the position and the typical duration of the sound objects, can help guiding the choice of the dictionary (harmonic molecules, chirplets, atoms with predetermined positions, ...).

Conversely, in other contexts, it can be desirable to build the dictionary with data-driven approaches, i.e. training examples of signals belonging to the same class (for example, the same speaker or the same musical instrument, ...). In this respect, Principal Component Analysis (PCA) offers interesting properties, but other approaches can be considered (in particular the direct optimization of the sparsity of the decomposition, or properties on the approximation error with M terms) depending on the targeted application.

In some cases, the training of the dictionary can require stochastic optimization, but one can also be interested in EM-like approaches when it is possible to formulate the redundant representation approach within a probabilistic framework.

Extension of the techniques of adaptive representation can also be envisaged by the generalization of the approach to probabilistic dictionaries, i.e. comprising vectors which are random variables rather than deterministic signals. Within this framework, the signal $y(t)$ is modeled as the linear combination of observations emitted by each element of the dictionary, which makes it possible to gather in the same model several variants of the same sound (for example various waveforms for a noise, if they are equivalent for the ear). Progress in this direction are conditioned to the definition of a realistic generative model for the elements of the dictionary and the development of effective techniques for estimating the model parameters.

3.3.5. Compressive sensing

The theoretical results around sparse representations have laid the foundations for a new research field called compressed sensing, emerging primarily in the USA. Compressed sensing investigates ways in which we can sample signals at roughly the lower information rate rather than the standard Shannon-Nyquist rate for sampled signals.

In a nutshell, the principle of Compressed Sensing is, at the acquisition step, to use as samples a number of random linear projections. Provided that the underlying phenomenon under study is sufficiently sparse, it is possible to recover it with good precision using only a few of the random samples. In a way, Compressed Sensing can be seen as a generalized sampling theory, where one is able to trade bandwidth (i.e. number of samples) with computational power. There are a number of cases where the latter is becoming much more accessible than the former; this may therefore result in a significant overall gain, in terms of cost, reliability, and/or precision.

MIMETIC Team

3. Scientific Foundations

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.

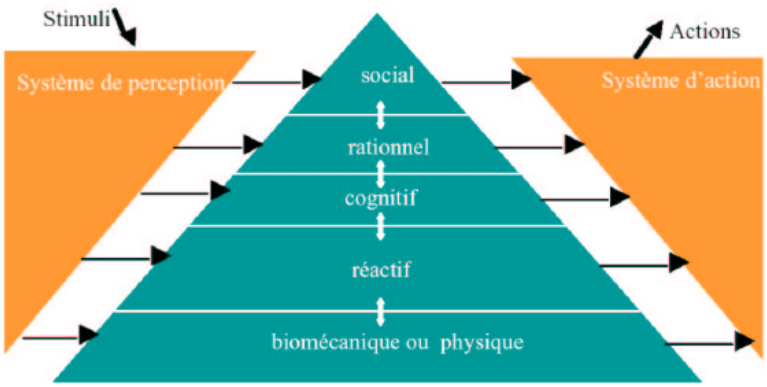


Figure 3. Layers of the motion control natural system in humans.

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 exists 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 propose models compatible with performance constraints imposed by interactive applications.

MINT Project-Team

3. Scientific Foundations

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 [30]. 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 [35]. Rapid and iterative prototyping techniques allow to decide between alternative solutions [31]. Controlled studies based on experimental or quasi-experimental designs can then be used to evaluate the chosen solutions [37]. 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 [34]. 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 fundations 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 [38], Dynamic Time Warping [40] 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 [36]; 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 [39].
- *mapping to application*, that studies how to link gesture inputs to application. This ranges from transfer function that is classically involved in pointing tasks [32], 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 Team

3. Scientific Foundations

3.1. Shape Acquisition

Recovering shapes from images is a fundamental task in computer vision. Applications are numerous and include, in particular, 3D modeling applications and mixed reality applications where real shapes are mixed with virtual environments. The problem faced here is to recover shape information such as surfaces from image information. A tremendous research effort has been made in the past to solve this problem in the static case and a number of solutions had been proposed. However, a fundamental issue still to be addressed is the recovery of full shape models with possibly evolving topologies using time sequence information. The main difficulties are precision, robustness of computed shapes as well as consistency of these models over time. Additional difficulties include the integration of multi-modality sensors as well as real-time applications.

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. Spectral Geometry

Spectral geometry processing consists of designing methods to process and transform geometric objects that operate in frequency space. This is similar to what is done in signal processing and image processing where signals are transposed into an alternative frequency space. The main interest is that a 3D shape is mapped into a spectral space in a pose-independent way. In other words, if the deformations undergone by the shape are metric preserving, all the meshes are mapped to a similar place in spectral space. Recovering the coherence between shapes is then simplified, and the spectral space acts as a "common language" for all shapes that facilitates the computation of a one-to-one mapping between pairs of meshes and hence their comparisons. However, several difficulties arise when trying to develop a spectral processing framework. The main difficulty is to define a spectral function basis on a domain which is a 2D (resp. 3D for moving objects) manifold embedded in 3D (resp. 4D) space and thus has an arbitrary topology and a possibly complicated geometry.

3.4. Surface Deformation

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.

3.5. Manifold Learning

The goal of motion analysis is to understand the movement in terms of movement coordination and corresponding neuromotor and biomechanical principles. Most existing tools for motion analysis consider as input rotational parameters obtained through an articulated body model, e.g. a skeleton; such model being tracked using markers or estimated from shape information. Articulated motion is then traditionally represented by trajectories of rotational data, each rotation in space being associated to the orientation of one limb segment in the body model. This offers a high dimensional parameterization of all possible poses. Typically, using a standard set of articulated segments for a 3D skeleton, this parameterization offers a number of degrees of freedom (DOF) that ranges from 30 to 40. However, it is well known that for a given motion performance, the trajectories of these DOF span a much reduced space. Manifold learning techniques on rotational data have proven their relevance to represent various motions into subspaces of high-level parameters. However, rotational data encode motion information only, independently of morphology, thus hiding the influence of shapes over motion parameters. One of the objectives is to investigate how motions of human and animal bodies, i.e. dense surface data, span manifolds in higher dimensional spaces and how these manifolds can be characterized. The main motivation is to propose morpho-dynamic indices of motion that account for both shape and motion. Dimensionality reduction will be applied on these data and used to characterize the manifolds associated to human motions. To this purpose, the raw mesh structure cannot be statistically processed directly and appropriate features extraction as well as innovative multidimensional methods must be investigated.

MOSTRARE Project-Team

3. Scientific Foundations

3.1. Modeling XML document transformations

Participants: Guillaume Bagan, Adrien Boiret, Iovka Boneva, Angela Bonifati, Anne-Cécile Caron, Benoît Groz, Joachim Niehren, Yves Roos, Sławek Staworko, Sophie Tison, Antoine Ndione, Tom Sebastian.

XML document transformations can be defined in W3C standards languages XQuery or XSLT. Programming XML transformations in these languages is often difficult and error prone even if the schemata of input and output documents are known. Advanced programming experience and considerable programming time may be necessary, that are not available in Web services or similar scenarios.

Alternative programming language for defining XML transformations have been proposed by the programming language community, for instance XDuce [46], Xtatic [44], [49], and CDuce [35], [36], [37]. The type systems of these languages simplify the programming tasks considerably. But of course, they don't solve the general difficulty in programming XML transformations manually.

Languages for defining node selection queries arise as sub-language of all XML transformation languages. The W3C standards use XPath for defining monadic queries, while XDuce and CDuce rely on regular queries defined by regular pattern equivalent to tree automata. Indeed, it is natural to look at node selection as a simple form of tree transformation. Monadic node selection queries correspond to deterministic transformations that annotate all selected nodes positively and all others negatively. N-ary node selection queries become non-deterministic transformations, yielding trees annotated by Boolean vectors.

After extensive studies of node selection queries in trees (in XPath and many other languages) the XML community has started more recently to formally investigate XML tree transformations. The expressiveness and complexity of XQuery are studied in [48], [57]. Type preservation is another problem, i.e., whether all trees of the input type get transformed into the output type, or vice versa, whether the inverse image of the output type is contained in the input type [52], [50].

The automata community usually approaches tree transformations by tree transducers [42], i.e., tree automata producing output structure. Macro tree transducers, for instance, have been proposed recently for defining XML transformations [50]. From the view point of logic, tree transducers have been studied for MSO definability [43].

3.2. Machine learning for XML document transformations

Participants: Jean Decoster, Pascal Denis, Jean-Baptiste Faddoul, Rémi Gilleron, Grégoire Laurence, Aurélien Lemay, Joachim Niehren, Sławek Staworko, Marc Tommasi, Fabien Torre, Gemma Garriga, Antonino Freno, Thomas Ricatte, Mikaela Keller.

Automatic or semi-automatic tools for inferring tree transformations are needed for information extraction. Annotated examples may support the learning process. The learning target will be models of XML tree transformations specified in some of the languages discussed above.

Grammatical inference is commonly used to learn languages from examples and can be applied to learn transductions. Previous work on grammatical inference for transducers remains limited to the case of strings [38], [53]. For the tree case, so far only very basic tree transducers have been shown to be learnable, by previous work of the Mostrare project. These are node selecting tree transducer (NSTTs) which preserve the structure of trees while relabeling their nodes deterministically.

Statistical inference is most appropriate for dealing with uncertain or noisy data. It is generally useful for information extraction from textual data given that current text understanding tools are still very much limited. XML transformations with noisy input data typically arise in data integration tasks, as for instance when converting PDF into XML.

Stochastic tree transducers have been studied in the context of natural language processing [45], [47]. A set of pairs of input and output trees defines a relation that can be represented by a 2-tape automaton called a *stochastic finite-state transducer* (SFST). A major problem consists in estimating the parameters of such transducer. SFST training algorithms are lacking so far [41].

Probabilistic context free grammars (pCFGs) [51] are used in the context of PDF to XML conversion [39]. In the first step, a labeling procedure of leaves of the input document by labels of the output DTD is learned. In the second step, given a CFG as a generative model of output documents, probabilities are learned. Such two steps approaches are in competition with one step approaches estimating conditional probabilities directly.

A popular non generative model for information extraction is *conditional random fields* (CRF, see a survey [54]). One main advantage of CRF is to take into account long distance dependencies in the observed data. CRF have been defined for general graphs but have mainly been applied to sequences, thus CRF for XML trees should be investigated.

So called *structured output* has recently become a research topic in machine learning [56], [55]. It aims at extending the classical categorization task, which consists to associate one or some labels to each input example, in order to handle structured output labels such as trees. Applicability of structured output learning algorithms remains to be asserted for real tasks such as XML transformations.

OAK Team

3. Scientific Foundations

3.1. Efficient XML and RDF data management

The development of Web technologies has led to a strong increase in the number and complexity of the applications which represent their data in Web formats, among which XML (for structured documents) and RDF (for Semantic Web data) are the most prominent. Oak has carried on research on algorithms and systems for efficiently processing expressive queries on such Web data formats. We have considered the efficient management of XML and RDF data, both for query evaluation and for efficiently applying updates, possibly in concurrence with queries. We have also started investigating multidimensional data analysis within RDF data warehouses.

For applications that integrate such Web data from various sources, we developed efficient and effective techniques to automatically recognize multiple representations of the same real-world object. That is, we devised main-memory resident algorithms that apply on hierarchical data [8] as well as algorithms that manipulate graph data leveraging off-the-shelf database management systems and parallelization to address both efficiency and scalability beyond main-memory [7].

3.2. Cloud-based Data Management

We have recently started to work on the efficient management of complex Web data, in particular structured XML documents and Semantic Web data under the form of RDF, in a cloud-based data management platform. We have investigated architectures and algorithms for storing Web data in elastic cloud-based stores and building an index for such data within the efficient key-value stores provided by off-the-shelf cloud data management platform. We have devised and prototyped such platforms for both XML and RDF data, and started experimenting with them in the Amazon Web Services (AWS) platform [13], [12], [10].

3.3. Data Transformation Management

With the increasing complexity of data processing 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. This includes systematically verifying, maintaining, and testing the transformations an application relies on. 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 query. To this end, we have investigated how to leverage data provenance (the information of the origin of data and the query operators) for query debugging. More specifically, we developed and implemented algorithms to explain unexpected results produced by a query (why-provenance) as well as expected results that are however missing from the query result (why-not provenance). Results have been presented in form of a software prototype [15].

ORPAILLEUR Project-Team

3. Scientific Foundations

3.1. From KDD to KDDK

knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining methods

Knowledge discovery in databases is a process for extracting knowledge units from large databases, units that can be interpreted and reused within knowledge-based systems.

Knowledge discovery in databases is a process for extracting knowledge units from large databases, units that can be interpreted and reused within knowledge-based systems. From an operational point of view, the KDD process is performed within a KDD system including 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 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, and concept lattice design (Formal Concept Analysis and extensions [95], [108]).
- Numerical methods are based on second-order Hidden Markov Models (HMM2, designed for pattern recognition [107]). Hidden Markov Models have good capabilities for locating stationary segments, and are mainly used for mining temporal and spatial data.

The principle summarizing KDDK can be understood as a process going from complex data units to knowledge units being guided by domain knowledge (KDDK or "knowledge with/for knowledge") [104]. Two original aspects can be underlined: (i) the KDD process is guided by domain knowledge, and (ii) the extracted units are embedded within a knowledge representation formalism to be reused in a knowledge-based system for problem solving purposes.

The various instantiations of the KDDK process in the research work of Orpailleur are mainly based on *classification*, considered as a polymorphic process involved in tasks such as modeling, mining, representing, and reasoning. Accordingly, the KDDK process may feed knowledge-based systems to be used for problem-solving activities in application domains, e.g. agronomy, astronomy, biology, chemistry, and medicine, and also for semantic web activities involving text mining, information retrieval, and ontology engineering [97], [81].

3.2. Methods for Knowledge Discovery guided by Domain Knowledge

knowledge discovery in databases guided by domain knowledge, lattice-based classification, formal concept analysis, frequent itemset search, association rule extraction, second-order Hidden Markov Models, stochastic process, numerical data mining method

knowledge discovery in databases guided by domain knowledge is a KDD process guided by domain knowledge ; the extracted units are represented within a knowledge representation formalism and embedded within a knowledge-based system.

Classification problems can be formalized by means of a class of individuals (or objects), a class of properties (or attributes), 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 (from a binary database) a set of formal concepts organized within a concept lattice [95]. Concept lattices are sometimes also called Galois lattices [82].

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 may 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 [123], [122].

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 [124], [125].

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. These 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. A special research effort focuses on the combination of knowledge elicited by experts and time-space regularities as extracted by an unsupervised classification based on stochastic models [23].

3.3. Elements on Text Mining

knowledge discovery form large collection of texts, text mining, information extraction, document annotation, ontologies

Text mining is a process for extracting knowledge units from large collections of texts, units that can be interpreted and reused within knowledge-based systems.

The objective of a text mining process is to extract new and useful knowledge units in a large set of texts [80], [89]. 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 the mining process more complex. 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, the text mining process is aimed at extracting new knowledge units from texts with the help of background knowledge encoded within an ontology and which is useful to relate notions present in a text, to guide and to help the text mining process. Text mining is especially useful in the context of semantic web for ontology engineering [86], [85], [84]. In the Orpailleur team, the focus is put on real-world texts in application domains such as astronomy, biology and medicine, using mainly symbolic data mining methods. 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. Elements on Knowledge Systems and Semantic Web

knowledge representation, ontology, description logics, classification-based reasoning, case-based reasoning, semantic web, knowledge-based information retrieval, web mining

Knowledge representation is a process for representing knowledge within an ontology using a knowledge representation formalism, giving knowledge units a syntax and a semantics. Semantic web is based on ontologies and allows search, manipulation, and dissemination of documents on the web by taking into account their contents, i.e. the semantics of the elements included in the documents.

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 very difficult and tedious. Semantic web is an attempt for guiding search for information with the help of machines, that are in charge of asking questions, searching for answers, classifying and interpreting the answers. However, a machine 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. This is why ontologies are of main importance with respect to the task of setting up semantic web. Thus, there is a need for representation languages for annotating documents, i.e. describing the content of documents, and giving a semantics to this content. Knowledge representation languages are (the?) good candidates for achieving the task: they have a syntax with an associated semantics, and they can be used for retrieving information, answering queries, and reasoning.

Semantic web constitutes a good platform for experimenting ideas on knowledge representation, reasoning, and KDDK. In particular, the knowledge representation language used for designing ontologies is the OWL language, which is based on description logics (or DL [79]). 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 that is 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 associated to case-based reasoning (CBR), that 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.

In the trend of semantic web, research work is also carried on semantic wikis which are wikis i.e., web sites for collaborative editing, in which documents can be annotated thanks to semantic annotations and typed relations between wiki pages. Such links provide kind of primitive knowledge units that can be used for guiding information retrieval or knowledge discovery.

PAROLE Project-Team

3. Scientific Foundations

3.1. Introduction

Research in speech processing gave rise to two kinds of approaches:

- research that aims at explaining how speech is produced and perceived, and that therefore includes physiological aspects (vocal tract control), physical (speech acoustics), psychoacoustics (peripheral auditory system), and cognitive aspects (building sentences),
- research aiming at modeling the observation of speech phenomena (spectral analysis, stochastic acoustic or linguistic models).

The former research topic is motivated by the high specificity of speech among other acoustical signals: the speech production system is easily accessible and measurable (at least at first approach); acoustical equations are reasonably difficult from a mathematical point of view (with simplifications that are moderately restrictive); sentences built by speakers are governed by vocabulary and grammar of the considered language. This led acousticians to develop research aiming at generating artificial speech signals of good quality, and phoneticians to develop research aiming at finding out the origin of speech sound variability and at explaining how articulators are utilized, how sounds of a language are structured and how they influence each other in continuous speech. Lastly, that led linguists to study how sentences are built. Clearly, this approach gives rise to a number of exchanges between theory and experimentation and it turns out that all these aspects of speech cannot be mastered easily at the same time.

Results available on speech production and perception do not enable using an analysis by synthesis approach for automatic speech recognition. Automatic speech recognition thus gives rise to a second approach that consists in modeling observations of speech production and perception. Efforts focused onto the design of numerical models (first simple vectors of spectral shapes and now stochastic or neural models) of word or phoneme acoustical realizations, and onto the development of statistical language models.

These two approaches are complementary; the latter borrows theoretical results on speech from the former, which, in its turn, borrows some numerical methods. Spectral analysis methods are undoubtedly the domain where exchanges are most marked. The simultaneous existence of these two approaches is one of the particularities of speech research conducted in Nancy and we intend to enhance exchanges between them. These exchanges will probably grow in number because of new applications like: **(i)** computer aided foreign language learning which requires both reliable automatic speech recognition and fine acoustic and articulatory speech analysis, **(ii)** automatic recognition of spontaneous speech which requires robustness against noise and speaker variability.

3.2. Speech Analysis and Synthesis

Our research activities focus on acoustical and perceptual cues of speech sounds, speech modifications and acoustic-to-articulatory inversion. Our main applications concern the improvement of the oral component of language learning, speech synthesis and esophageal voices.

3.2.1. Oral comprehension

We developed tools to improve speech perception and production, and made perceptual experiments to prove their efficiency in language learning. These tools are also of interest for hearing impaired people, as well as for normally hearing people in noisy environments and also for children who learn to read (children who have language disabilities without cognitive deficit or hearing impairment and "normal" children).

3.2.1.1. Computer-assisted learning of prosody

We are studying automatic detection and correction of prosodic deviations made by a learner of a foreign language. This work implies three different tasks: (a) the detection of the prosodic entities of the learner's realization (lexical accent, intonative patterns), (b) the evaluation of the deviations, by comparison with a model, and (c) their corrections, both verbal and acoustic. This last kind of feedback is directly done on the learner's realization: the deviant prosodic cues are replaced by the prosodic cues of the reference. The identification and correction tasks use speech analysis and modification tools developed in our team.

Within the framework of a new project (see 6.1.6.2), we also investigate the impact of a language intonational characteristics on the perception and production of the intonation of a foreign language.

3.2.1.2. Phonemic discrimination in language acquisition and language disabilities

We keep working on a project concerning identification of early predictors of reading, reading acquisition and language difficulties, more precisely in the field of specific developmental disabilities : dyslexia and dysphasia. A fair proportion of those children show a weakness in phonological skills, particularly in phonemic discrimination. However, the precise nature and the origin of the phonological deficits remain unspecified. In the field of dyslexia and normal acquisition of reading, our first goal was to contribute to identify early indicators of the future reading level of children.We based our work on the longitudinal study - with 85 French children - of [45], [46] which indicates that phonemic discrimination at the beginning of kindergarten is strongly linked to success and specific failure in reading acquisition. We study now the link between oral discrimination both with oral comprehension and written compréhension. Our analyses are based on the follow up of a hundred chrildren for 4 years from kindergarten to end of grade 2 (from age 4 to age 8). Publications in progress.

3.2.1.3. Esophageal voices

It is possible for laryngectomees to learn a substitution voice: the esophageal voice. This voice is far from being natural. It is characterized by a weak intensity, a background noise that bothers listening, and a low pitch frequency. A device that would convert an esophageal voice to a natural voice would be very useful for laryngectomees because it would be possible for them to communicate more easily. Such natural voice restitution techniques would ideally be implemented in a portable device.

3.2.2. Acoustic-to-articulatory inversion

Acoustic-to-articulatory inversion aims at recovering the articulatory dynamics from speech signal that may be supplemented by images of the speaker face. Potential applications concern low bit rate speech coding, automatic speech recognition, speech production disorders assessment, articulatory investigations of phonetics, talking heads and articulatory feedback for language acquisition or learning.

Works on acoustic-to-articulatory inversion widely rely on an analysis by synthesis approach that covers three essential aspects:

Solving acoustic equations. In order to solve the acoustic equations adapted to the vocal tract, one assumes that the sound wave is a plane wave in the vocal tract and that it can be unbend. There are two families of solving methods:

(i) frequency methods through the acoustical-electrical analogy,

(ii) spatio-temporal methods, through the direct solving of finite difference equations derived from Webster equations.

Measuring the vocal tract. This represents an important obstacle because there does not exist any reliable method enabling a precise measurement in time and dimension. MRI (Magnetic Resonance Imaging) enables 3D measurements but is not sufficiently fast and X-rays only allows a sagittal slice of the vocal tract to be captured while involving not acceptable health hazards.

Articulatory modeling. Articulatory models aim at describing all the possible vocal tract shapes with a small number of parameters, while preserving deformations observed on a real vocal tract. Present articulatory models often derive from data analysis of cineradiography moving pictures. One of the most widely used is the one built by Maeda [52].

One of the major difficulties of inversion is that an infinity of vocal tract shapes can give rise to the same speech spectrum. Acoustic-to-articulatory inversion methods are categorized into two families:

- methods that optimize a function generally combining speaker's articulatory effort and acoustical distance between natural and synthesized speech. They exploit constraints allowing the number of possible vocal tract shapes to be reduced.
- table look-up methods resting on an articulatory codebook of articulatory shapes indexed by their acoustical parameters (generally formant frequencies). After possible shapes have been recovered at each time, an optimization procedure is used to find an inverse solution in the form of an optimal articulatory path.

As our contribution only concerns inversion, we accepted widely used articulatory synthesis methods. We therefore chose Maeda's articulatory model, the acoustical-electrical analogy to compute the speech spectrum and the spatio-temporal method proposed by Maeda to generate the speech signal. As regards inversion, we chose Maeda's model to constrain vocal tract shapes because this model guarantees that synergy and compensation articulatory phenomena are still possible, and consequently, that articulatory deformations close to those of a human speaker may be recovered. The most important challenges in this domain are the inversion of any class of speech sounds and to perform inversion from standard spectral data, MFCC for instance. Indeed at present, only vowels and sequences of vowels can be inverted, and only some attempts concern fricatives sounds. Moreover, most of the inversion techniques use formant frequencies as input data although formants cannot be extracted from speech easily and reliably.

3.2.3. Strategies of labial coarticulation

The investigation of labial coarticulations strategies is a crucial objective with the view of developing a talking head which would be understandable by lip readers, especially deaf persons.

In the long term, our goal is to determine a method of prediction of labial coarticulation adaptable to a virtual speaker. Predicting labial coarticulation is a difficult problem that gave rise to many studies and models. To predict the anticipatory coarticulation gestures (see [42] for an overall presentation of labial coarticulation), three main models have been proposed: the look-ahead model, the time-locked model and the hybrid model.

These models were often compared on their performance in the case of the prediction of anticipation protrusion in VCV or VCCV sequences where the first vowel is unrounded, the consonant(s) is neutral with respect to labial articulation and the last vowel is rounded. There is no general agreement about the efficiency of these models. More recent models have been developed. The one of Abry and Lallouache [37] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [40] proposed dominance functions that require a substantial numerical training.

Most of these models derive from the observations of a limited number of speakers. We are thus developing a more explicative model, i.e., essentially a phonetically based approach that tries to understand how speakers manage to control labial parameters from the sequence of phonemes to be articulated.

3.2.4. Speech Synthesis

Data-driven speech synthesis is widely adopted to develop Text-to-Speech (TTS) synthesis systems. Basically, it consists of concatenating pieces of signal (units) selected from a pre-recorded sentence corpus. Our ongoing work on acoustic TTS was recently extended to study acoustic-visual speech synthesis (bimodal units).

3.2.4.1. Text-to-speech synthesis

Data-driven text-to-speech synthesis is usually composed of three steps to transform a text in speech signal. The first step is Natural Language Processing (NLP) which tags and analyzes the input text to obtain a set of features (phoneme sequence, word grammar categories, syllables...). It ends with a prosodic model which transforms these features into acoustic or symbolic features (F0, intensity, tones...). The second step uses a Viterbi algorithm to select units from a corpus recorded beforehand, which have the closest features to the prosodic features expected. The last step amounts to concatenate these units.

Such systems usually generate a speech signal with a high intelligibility and a naturalness far better than that achieved by old systems. However, building such a system is not an easy task [39] and the global quality mainly relies on the quality of the corpus and prosodic model. The prosodic model generally provides a good standard prosody, but, the generated speech can suffer from a lack of variability. Especially during the synthesis of extended passages, repetition of similar prosodic patterns can lead to a monotonous effect. Therefore, to avoid this problem due to the projection of linguistic features onto symbolic or acoustic dimensions (during NLP), we [41] proposed to perform the unit selection directly from linguistic features without incorporating any prosodic information. To compensate the lack of prosodic prediction, the selection needs to be performed with numerous linguistic features. The selection is no longer restrained by a prosodic model but only driven by weighted features. The consequence is that the quality of synthesis may drop in crucial instants. Our works deal to overcome this new problem while keeping advantage of the lack of prosodic model.

These works have an impact on the construction of corpus and on the NLP engine which needs to provide as much information as possible to the selection step. For instance, we introduced a chunker (shallow parser) to give us information on a potential rhythmic structure. Moreover, to perform the selection, we developed an algorithm to automatically weight the linguistic features given by the NLP. Our method relies on acoustic clustering and entropy information [41]. The originality of our approach leads us to design a more flexible unit selection step, constrained but not restrained.

3.2.4.2. Acoustic-visual speech synthesis

Audiovisual speech synthesis can be achieved using 3D features of the human face supervised by a model of speech articulation and face animation. Coarticulation is approximated by numerical models that describe the synergy of the different articulators. Acoustic signal is usually synthetic or natural speech synchronized with the animation of the face. Some of the audiovisual speech systems are inspired by recent development in speech synthesis based on samples and concatenative techniques. The main idea is to concatenate segments of recorded speech data to produce new segments. Data can be video or motion capture. The main drawback of these methods is that they focus on one field, either acoustic or visual. But (acoustic) speech is actually generated by moving articulators, which modify the speaker's face. Thus, it is natural to find out that acoustic and face movements are correlated. A key point is therefore to guarantee the internal consistency of the acoustic-visual signal so that the redundancy of these two signals acknowledged as a determining perceptive factor, can really be exploited by listeners. It is thus important to deal with the two signals (acoustic and visual) simultaneously and to keep this link during the whole process. This is why we make the distinction between audiovisual speech synthesis (where acoustic is simply synchronized with animation) and acoustic-visual speech where speech is considered as a bimodal signal (acoustic and visual) as considered in our work. Our long-term goal is to contribute to the fields of acoustic speech synthesis and audiovisual speech synthesis by building a bimodal corpus and developing an acoustic-visual speech synthesis system using bimodal unit concatenation.

3.3. Automatic speech recognition

Automatic speech recognition aims at reproducing the cognitive ability of humans to recognize and understand oral speech. Our team has been working on automatic speech recognition for decades. We began with knowledge-based recognition systems and progressively made our research works evolve towards stochastic approaches, both for acoustic and language models. Regarding acoustic models, we have especially investigated HMM (Hidden Markov Models), STM (Stochastic Trajectory Models), multi-band approach and BN (Bayesian Networks). Regarding language models, our main interest has concerned ngram approaches (word classes, trigger, impossible ngram, etc).

The main challenge of automatic speech recognition is its robustness to multiple sources of speech variability [44]. Among them, we have been focusing on acoustic environment, inter- and intra-speaker variability, different speaking styles (prepared speech, spontaneous, etc.) and non-native pronunciations.

Another specifity of automatic speech recognition is the necessity to combine efficiently all the research works (in acoustic modeling, langage modeling, speaker adaptation, etc.) into a core platform in order to evaluate them, and to go beyond pure textual transcriptions by enriching them with punctuation, syntax, etc., in order to make them exploitable by both humans and machines.

3.3.1. Acoustic features and models

The raw acoustic signal needs to be parameterized to extract the speech information it contains and to reduce its dimensionality. Most of our research and recognition technologies make use of the classical Mel Feature Cepstral Coefficients, which have proven since many years to be amongst the most efficient front-end for speech recognition. However, we have also explored alternative parameterizations to support some of our recent research progresses. For example, prosodic features such as intonation curves and vocal energy give important cues to recognize dialog acts, and more generally to compute information that relates to supra-phonemic (linguistic, dialog, ...) characteristics of speech. Prosodic features are developed jointly for both the Speech Analysis and Speech Recognition topics. We also developed a new robust front-end, which is based on wavelet-decomposition of the speech signal.

Concerning acoustic models, stochastic models are now the most popular approach for automatic speech recognition. Our research on speech recognition also largely exploits Hidden Markov Models (HMM). In fact, HMMs are mainly used to model the acoustic units to be recognized (usually triphones) in all of our recognition engines (ESPERE, ANTS...). Besides,we have investigated Bayesian Networks (BN) to explicitly represent random variables and their independence relationships to improve noise robustness.

3.3.2. Robustness and invariance

Part of our research activities about ASR aims at improving the robustness of recognizers to the different sources of variability that affect the speech signal and damage the recognition. Indeed, the issue of the lack of robustness of state-of-the-art ASR systems is certainly the most problematic one that still prevents the wide deployment of speech recognizers nowadays. In the past, we developed a large range of techniques to address this difficult topic, including robust acoustic models (such as stochastic trajectory and multi-band models) and model adaptation techniques (such as missing data theory). The following state-of-the-art approaches thus form our baseline set of technologies: MLLR (Maximum Likelihood Linear Regression), MAP (Maximum A Posteriori), PMC (Parallel Model Combination), CMN (Cepstral Mean Normalization), SAT (Speaker Adaptive Training), HLDA (Heteroscedastic Linear Discriminant Analysis), Spectral Subtraction and Jacobian Adaptation.

These technologies constitute the foundations of our recent developments in this area, such as non-native speaker adaptation, out-of-vocabulary words detection and adaptation to pronunciation variations. Handling speech variabilities may also benefit from exploiting additional external or contextual sources of information to more tightly guide the speech decoding process. This is typically the role of the language model, which shall in this context be augmented with higher-level knowledge, such as syntactic or semantic cues. Yet, automatically extracting such advanced features is very challenging, especially on imperfect transcribed speech.

The performance of automatic speech recognition (ASR) systems drastically drops when confronted with non-native speech. If we want to build an ASR system that takes into account non-native speech, we need to modify the system because, usually, ASR systems are trained on standard phone pronunciations and designed to recognize only native speech. In this way, three method categories can be applied: acoustic model transformation, pronunciation modeling and language modeling. Our contribution concerns the first two methods.

3.3.3. Segmentation

Audio indexing and automatic broadcast news transcription need the segmentation of the audio signal. The segmentation task consists in two steps: firstly, homogeneous segments are extracted and classified into speech, noise or music, secondly, speakers turns are detected in the extracted speech segments.

Speech/music segmentation requires to extract discriminant acoustic parameters. Our contribution concerns the MFCC and wavelet parameters. Another point is to find a good classifier. Various classifiers are commonly used: k-Nearest-Neighbors, Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks.

As to detect speaker turns, the main approach consists of splitting the audio signal into segments that are assumed to contain only one speaker and then a hierarchical clustering scheme is performed for merging segments belonging to the same speaker.

3.3.4. Speech/text alignment

Speech/text alignment consists in finding time boundaries of words or phones in the audio signal knowing the orthographic transcription. The main applications of speech/text alignment are training of acoustic models, segmentation of audio corpus for building units for speech synthesis or segmentation of the sentence uttered by a learner of a foreign language. Moreover, speech/text alignement is a useful tool for linguistic researchers.

Speech/text alignment requires two steps. The first step generates the potential pronunciations of the sentence dealing with multiple pronunciations of proper nouns, liaisons, phone deletions, and assimilations. For that, the phonetizer is based on a phonetic lexicon, and either phonological rules or an automatic classifier as a decision tree. The second step finds the best pronunciation corresponding to the audio signal using acoustic HMM models and an alignment algorithm. The speech team has been working on this domain for a long time.

3.4. Speech to Speech Translation and Langage Modeling

Speech-to-Speech Translation aims at translating a source speech signal into a target speech signal. A sequential way to adress this problem is to first translate a text to another one. And after, we can connect a speech recognition system at the input and a text to speech synthesis system at the output. Several ways to adress this issue exist. The concept used in our group is to let the computer learning from a parallel text all the associations between source and target units. A unit could be a word or a phrase. In the early 1990s [38] proposes five statistical translation models which became inescapable in our community. The basic idea of the model 1 is to consider that any word of the target language could be a potential translation of any source word. The problem is then to estimate the distribution probability of a target word given a source one. The translation problem is similar to the speech recognition one. Indeed, we have to seek the best foreign sentence given a source one. This one is obtained by decoding a lattice translation in which a language and translation models are used. Several issues have to be supported in machine translation as described below.

3.4.1. Word translation

The first translation systems identify one-to-one associations between words of target and source languages. This is still necessary in the present machine translation systems. In our group we develop a new concept to learn the translation table. This approach is based on computing all the inter-lingual triggers inside a parallel corpus. This leads to a pertinent translation table [51]. Obviously, this is not sufficient in order to make a realistic translation because, with this approach, one word is always translated into one word. In fact, it is possible to express the same idea in two languages by using different numbers of words. Thus, a more general one-to-one alignement has to be achieved.

3.4.2. Phrase translation

The human translation is a very complex process which is not only word-based. A number of research groups developed phrase-based systems which are different from the baseline IBM's model in training. These methods, deals with linguistic units which consists in more than one word. The model supporting phrase-based machine translation uses reordering concept and additional feature functions. In order to retrieve phrases, several approaches have been proposed in the litterature. Most of them require word-based alignments. For example, Och and al. [54] collected all phrase pairs that were consistent with the word alignment provided by Brown's models.

We developed a phrase based algorithm which is based on finding first an adequate list of phrases. Then, we find out the best corresponding translations by using our concept of inter-lingual triggers. A list of the best translations of sequences is then selected by using simulated annealing algorithm.

3.4.3. Language model

A language model has an important role in a statistical machine translation. It ensures that the translated words constitute a valid linguistic sentence. Most of the community uses n-grams models, that is what we do also.

3.4.4. Decoding

The translation issue is treated as an optimization problem. Translating a sentence from English into a Foreign language involves finding the best Foreign target sentence f^* which maximizes the probability of f given the English source sentence e. The Bayes rule allows to formulate the probability $P(f|e)$ as follows:

$$f^* = \arg\max_f P(f|e) = \arg\max_f P(e|f)P(f)$$

The international community uses either PHARAOH [48] or MOSES [47] based on a beam search algorithm. In our group we started decoding by PHARAOH but we moved recently to MOSES.

PERCEPTION Team

3. Scientific Foundations

3.1. The geometry of multiple images

Computer vision requires models that describe the image creation process. An important part (besides e.g. radiometric effects), concerns the geometrical relations between the scene, cameras and the captured images, commonly subsumed under the term “multi-view geometry”. This describes how a scene is projected onto an image, and how different images of the same scene are related to one another. Many concepts are developed and expressed using the tool of projective geometry. As for numerical estimation, e.g. structure and motion calculations, geometric concepts are expressed algebraically. Geometric relations between different views can for example be represented by so-called matching tensors (fundamental matrix, trifocal tensors, ...). These tools and others allow to devise the theory and algorithms for the general task of computing scene structure and camera motion, and especially how to perform this task using various kinds of geometrical information: matches of geometrical primitives in different images, constraints on the structure of the scene or on the intrinsic characteristics or the motion of cameras, etc.

3.2. The photometry component

In addition to the geometry (of scene and cameras), the way an image looks like depends on many factors, including illumination, and reflectance properties of objects. The reflectance, or “appearance”, is the set of laws and properties which govern the radiance of the surfaces . This last component makes the connections between the others. Often, the “appearance” of objects is modeled in image space, e.g. by fitting statistical models, texture models, deformable appearance models (...) to a set of images, or by simply adopting images as texture maps.

Image-based modelling of 3D shape, appearance, and illumination is based on prior information and measures for the coherence between acquired images (data), and acquired images and those predicted by the estimated model. This may also include the aspect of temporal coherence, which becomes important if scenes with deformable or articulated objects are considered.

Taking into account changes in image appearance of objects is important for many computer vision tasks since they significantly affect the performances of the algorithms. In particular, this is crucial for feature extraction, feature matching/tracking, object tracking, 3D modelling, object recognition etc.

3.3. Shape Acquisition

Recovering shapes from images is a fundamental task in computer vision. Applications are numerous and include, in particular, 3D modeling applications and mixed reality applications where real shapes are mixed with virtual environments. The problem faced here is to recover shape information such as surfaces, point positions, or differential properties from image information. A tremendous research effort has been made in the past to solve this problem and a number of partial solutions had been proposed. However, a fundamental issue still to be addressed is the recovery of full shape information over time sequences. The main difficulties are precision, robustness of computed shapes as well as consistency of these shapes over time. An additional difficulty raised by real-time applications is complexity. Such applications are today feasible but often require powerful computation units such as PC clusters. Thus, significant efforts must also be devoted to switch from traditional single-PC units to modern computation architectures.

3.4. Motion Analysis

The perception of motion is one of the major goals in computer vision with a wide range of promising applications. A prerequisite for motion analysis is motion modelling. Motion models span from rigid motion to complex articulated and/or deformable motion. Deformable objects form an interesting case because the models are closely related to the underlying physical phenomena. In the recent past, robust methods were developed for analysing rigid motion. This can be done either in image space or in 3D space. Image-space analysis is appealing and it requires sophisticated non-linear minimization methods and a probabilistic framework. An intrinsic difficulty with methods based on 2D data is the ambiguity of associating a multiple degree of freedom 3D model with image contours, texture and optical flow. Methods using 3D data are more relevant with respect to our recent research investigations. 3D data are produced using stereo or a multiple-camera setup. These data (surface patches, meshes, voxels, etc.) are matched against an articulated object model (based on cylindrical parts, implicit surfaces, conical parts, and so forth). The matching is carried out within a probabilistic framework (pair-wise registration, unsupervised learning, maximum likelihood with missing data).

Challenging problems are the detection and segmentation of multiple moving objects and of complex articulated objects, such as human-body motion, body-part motion, etc. It is crucial to be able to detect motion cues and to interpret them in terms of moving parts, independently of a prior model. Another difficult problem is to track articulated motion over time and to estimate the motions associated with each individual degree of freedom.

3.5. Multiple-camera acquisition of visual data

Modern computer vision techniques and applications require the deployment of a large number of cameras linked to a powerful multi-PC computing platform. Therefore, such a system must fulfill the following requirements: The cameras must be synchronized up to the millisecond, the bandwidth associated with image transfer (from the sensor to the computer memory) must be large enough to allow the transmission of uncompressed images at video rates, and the computing units must be able to dynamically store the data and/to process them in real-time.

Current camera acquisition systems are all-digital ones. They are based on standard network communication protocols such as the IEEE 1394. Recent systems involve as well depth cameras that produce depth images, i.e. a depth information at each pixel. Popular technologies for this purpose include the Time of Flight Cameras (TOF cam) and structured light cameras, as in the very recent Microsoft's Kinect device.

3.6. Auditory and audio-visual scene analysis

For the last two years, PERCEPTION has started to investigate a new research topic, namely the analysis of auditory information and the fusion between auditory and visual data. In particular we are interested in analyzing the acoustic layout of a scene (how many sound sources are out there and where are they located? what is the semantic content of each auditory signal?) For that purpose we use microphones that are mounted onto a human-like head. This allows the extraction of several kinds of auditory cues, either based on the time difference of arrival or based on the fact that the head and the ears modify the spectral properties of the sounds perceived with the left and right microphones. Both the temporal and spectral binaural cues can be used to locate the most proeminent sound sources, and to separate the perceived signal into several sources. This is however an extremely difficult task because of the inherent ambiguity due resemblance of signals, and of the presence of acoustic noise and reverberations. The combination of visual and auditory data allows to solve the localization and separation tasks in a more robust way, provided that the two stimuli are available. One interesting yet unexplored topic is the development of hearing for robots, such as the role of head and body motions in the perception of sounds.

POTIOC Team

3. Scientific Foundations

3.1. Introduction

The design of new user interfaces is a complex process that requires tackling research challenges at different levels. First, at a technological level, the input and output interaction space is becoming richer and richer. We will explore the new input/output modalities offered by such a technological evolution, and we will contribute to extend these modalities for the purpose of our main objective, which is to make 3D digital worlds available to all. Then, we will concentrate on the design of good interaction techniques that rely on such input/output modalities, and that are dedicated to the population targeted by this project, i.e. general public, specialists which are not 3D experts, and people with impairments. Finally, a large part of our work will be dedicated to the understanding and the assessment of user interaction. In particular, we will conduct user studies to guide the design of hardware and software UI, to evaluate them, and to better understand how a user interacts with 3D environments.

These three levels, input/output modalities, interaction techniques, and human factors will be the three main research directions of Potioc. Of course, they are extremely linked, and they cannot be studied independently, one after the other. In particular, user studies will follow the design process of hardware/software user interfaces from the beginning to the end, and both hardware and software exploration will be interdependent. The design of a new 3D user interface will thus require some work at different levels, as illustrated in Figure 2 . All members of Potioc will contribute in each of these research directions.

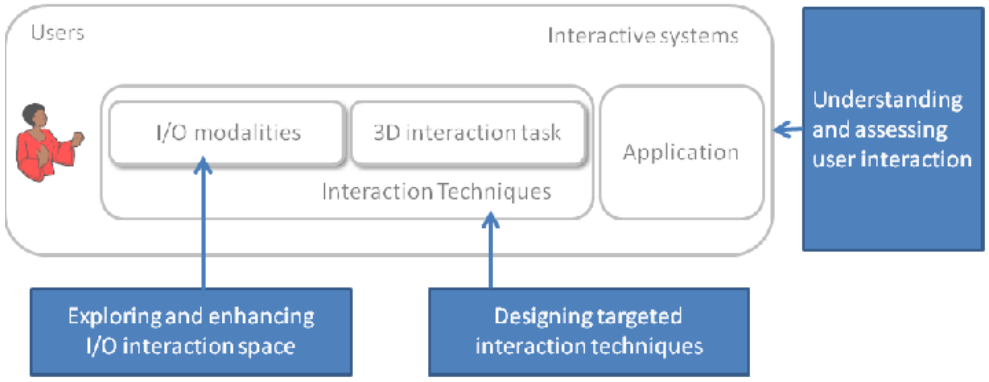


Figure 2. Diagram of an interactive system and the three main research axes of the Potioc project (blue boxes).

3.2. Exploring and enhancing input/output interaction space

The Potioc project-team will be widely oriented towards new innovative input and output modalities, even if standard approaches based on keyboard/mouse and standard screens will not be excluded. This includes motor-based interfaces, and physiological interfaces like BCI, as well as stereoscopic display and augmented reality setups. These technologies may have a great potential for opening 3D digital worlds to anyone, if they are correctly exploited.

We will explore various input/output modalities. Of course, we will not explore all of them at the same time, but we do not want to set an agenda either, for focusing on one of them. For a given need fed by end-users, we will choose among the various input/output modalities the ones that have the biggest potential. In the following paragraphs, we explain in more details the research challenges we will focus on to benefit from the existing and upcoming technologies.

3.2.1. Real-time acquisition and signal processing

There is a wide number of sensors that can detect users' activity. Beyond the mouse that detects x and y movements in a plane, various sensors are dedicated to the detection of 3D movements, pressure, brain and physiological activity, and so on. These sensors provide information that may be very rich, either to detect command intent from the user or to estimate and understand the user's state in real-time, but that are difficutly exploitable as it. Hence, a major challenge here is to extract the relevant information from the noisy raw data provided by the sensor.

An example, and important research topic in Potioc, is in the analysis of brain signals for the design of BCI. Indeed, brain signals are usually measured by EEG, such EEG signals being very noisy, complex and non-stationary. Moreover, for BCI-based applications, they need to be processed and analyzed in real-time. Finally, EEG signals exhibit large inter-user differences and there are usually few examples of EEG signals available to tune the BCI to a given user (we cannot ask the user to perform thousands of time the same mental task just to collect examples). As such, appropriate signal processing algorithms must be designed in order to robustly identify EEG patterns reflecting the user's intention. The research challenges are thus to design algorithms with high performance (in terms of rate of correctly identify user's state) anytime, anywhere, that are fully automatic and with minimal or no calibration time. In other words, we must design BCI that are convenient, comfortable and efficient enough so that they can be accepted and used by the end-user. Indeed, most users, in particular healthy users in the general public are used to highly convenient and efficient input devices (e.g., a simple mouse) and would not easily tolerate systems with a lower performance. Achieving this would make BCI good enough to be usable outside laboratories, e.g., for video gamers or patients. This will also make BCI valuable and reliable evaluation tools, e.g., to understand users' state during a given task. To address these challenges, pattern recognition and machine learning techniques are often used in order to find the optimal signal processing parameters. Similar approaches may contribute to the analysis of signals coming from other input devices than BCI. An example is the exploitation of depth cameras, where we need to find relevant information from noisy signals. Other emerging technologies will require similar attention, where the goal will be to transform an unstructured raw signal into a set of higher level descriptors that can be used as input parameters for controlling interaction techniques.

3.2.2. Restitution and perceptive feedback

Similarly to the input side, the feedback provided to the user through various output modalities will be explored in Potioc. Beyond the standard screens that are commonly used, we will explore various displays. In particular, in the scope of visual restitution, we will notably focus on large screens and tables, mobile setups and projection on real objects, and stereoscopic visualization. The challenge here will be to conceive good visual metaphors dedicated to these unconventional output devices in order to maximize the attractiveness and the pleasure linked to the use of these technologies.

For example, we will investigate the use of stereoscopic displays for extending the current visualization approaches. Indeed, stereoscopic visualization has been little explored outside the complex VR setups dedicated to professional users. We believe that this modality may be very interesting for non-expert users, in wider contexts. To reach this goal, we will thus concentrate on new visual metaphors that benefit from stereoscopic visualization, and we will explore how, when, and where stereoscopy may be used.

Depending on the targeted interaction tasks, we may also investigate various additional output modalities such as tangible interaction, audio displays, and so on. In any case, our approach will be the same, which is understanding how new perceptive modalities may push the frontier of our current interactive systems.

3.2.3. Creation of new systems

In addition to the exploration and the exploitation of existing input and output modalities for enhancing interaction with 3D content, we may also contribute to extend the current input/output interaction space by building new interactive systems. This will be done by combining hardware components, or by collaborating with mechanics/electronics specialists.

3.3. Designing targeted interaction techniques

In the previous section, we focused on the input/output interaction space, which is closely related to hardware components. In this part, we focus on the design of interaction techniques, which we define here as the mean through which a user will complete an interaction task from a given interaction space. Even if this is naturally also linked to the underlying hardware components, the research conducted in this axis of the project will mainly concern software developments.

Similar to the input/output interaction space, the design of interaction techniques requires focusing on both the motor and the sensory components. In our 3D spatial context, thus the challenges will be to find good mappings between the available input and the DOF that need to be controlled in the 3D environment, and to provide relevant feedback to users so that they can understand well what they are doing.

The design of interaction techniques should be strongly guided by the targeted end-users. For example, a 3D UI dedicated to an expert user will not suit a novice user, and the converse is also true. In Potioc, where the final goal is to open 3D digital worlds to anyone, we will concentrate on the general public, specialists that are not 3D experts, and people with impairments.

3.3.1. General public

3D UIs have mainly been designed for professional use. For example, modeling tools require expertize to be used correctly and, consequently, they exclude the general public from the process of creating 3D content. Similarly, immersive technologies have been dedicated to professional users for a long time. Therefore, immersive 3D interaction techniques have generally been thought for trained users, and they may not fit well with a general public context. In Potioc, an important motivation will be to re-invent 3D UIs to adapt them to the general public. This motivation will guide us towards new approaches that have been little explored until now. In particular, to reach our objective, we will give a strong importance to the following criteria:

- Intuitiveness: a very short learning curve will be required.
- Enjoyability: this is needed to motivate novice users in the complex process of interaction with 3D content.
- Robustness: the UIs should support untrained users that may potentially interact with unpredictable actions.

In addition, we will keep connected with societal and technological factors surrounding the general public. For example, [multi]touch-screens have become very popular these past few years, and everyone tend to be familiar with a standard gesture vocabulary (e.g. pinch gestures and flicking gestures). We will rely on these commonly acquired *way-of-interact* to optimize the acceptability of the 3D UIs we will design. In this part of the project the challenge will be to conceive 3D UIs that offer a high degree of interactivity, while ensuring an easy access to technology, as well as a wide adherence.

3.3.2. Specialists

General public will be one of the main targets of Potioc for the design of 3D UIs. However, we do not exclude specialists, who have little experience with 3D interaction. These specialists can be for example artists, archaeologists, or architects. In any case, we are convinced that 3D digital worlds could benefit to such categories of users if we propose dedicated 3D UIs that allows them to better understand, communicate, or create, with their respective skills. Because such specialists will gain expertize while interacting with 3D content, it will be necessary to design 3D UIs that can adapt to their evolving level of expertize. In particular, the UIs should be easy to use and attractive enough to encourage new users. At the same time, they should provide advanced features that the specialist can discover while gaining expertize.

3.3.3. People with impairments

While the general public has been only scarcely considered as a potential target audience for 3D digital worlds, another category of users is even more neglected: people with impairments. Indeed, such people, in particular those with motor impairments, are unable to use classical input devices, since they have been designed for healthy users. People with motor impairment have to use dedicated input devices, adapted to their disabilities, such as a single switch. Since such input devices usually have much fewer degrees of freedom than classical devices, it is necessary to come up with appropriate interaction techniques in order to efficiently use this limited number of DOF to still enable the user to perform complex tasks in the 3D environment. In Potioc, our focus will be on the use of BCI to enable motor impaired users to interact with 3D environment for learning, creation and entertainment. Indeed, BCI enable a user to interact without any motor movement.

3.4. Understanding and assessing user interaction

The exploration of the input/output interaction space, and the design of new interaction techniques, are strongly linked with human factors, which will be the third research axis of the Potioc project. Indeed, to guide the developments described in the previous sections, we first need to well understand users' motor and cognitive skills for the completion of 3D interaction tasks. This will be explored thanks to *a-priori* experiments. In order to evaluate our hardware and software interfaces, we will conduct *a-posteriori* user studies. Finally, we will explore new approaches for a real-time cognitive analysis of the performance and the experience of a user interacting with a 3D environment.

The main challenge in this part of the project will be to design good experimental protocols that will allow us to finely analyze various parameters for improving our interfaces. In 2D, there exist many standard protocols and prediction laws for evaluating UIs (e.g. Fitts law and ISO 9241). This is not the case in 3D. Consequently, a special care must be taken when evaluating interaction in 3D spatial contexts.

In addition to the standard experiments we will conduct in our lab, we will conduct large scale experiments thanks to the strong collaboration we have with the center for the widespread of scientific culture, Cap Sciences (see Collaboration section). With such kind of experiments, we will be able to test hundreds of participants, with various ages, gender, or level of expertize that we will be able to track thanks to the Navinum system [3], and this during long period of time. A challenge for us will be to gain benefit from this wealth of information for the development of our 3D UIs.

3.4.1. A-priori user studies

Before designing 3D UIs, it is important to understand what a user is good at, and what may cause difficulties. This is true at a motor level, as well as a cognitive level. For example, are users able to coordinate the movements of several fingers on a touchscreen at the same time, or are they able to finely control the quantity of force applied on it while moving their hand? Similarly, are the users able to mentally predict a 3D rotation, and how many levels of depth are they able to distinguish when visualizing stereoscopic images? To answer these questions, we will conduct preliminary studies.

Our research in that direction will guide our developments for the other research axes described above. For example, it will be interesting to explore touch-based 3D UIs that take into account several level of force if we see that this parameter can be easily handled by users. On the other hand, if the results of a-priori tests show that this input cannot be easily controlled, then we will not push forward that direction.

The members of Potioc have already conducted such kinds of experiments, and we will continue our work in that direction. For some investigations, we will collaborate with psychologists and experts in cognitive sciences (see Collaborations section) to explore in more depth motor and cognitive human skills.

[3] Navinum is a system based on a RFID technology that is used to collect informations about the activity of the visitors in Cap Sciences. http://www.scribd.com/doc/55178878/Dossier-de-Presse-Numerique-100511

A-priori studies will allow us to understand how users tend to "*naturally*" interact to complete 3D interaction tasks, and to understand which feedback are the best suited. This will be a first answer to our global quest of providing pleasant interfaces. Indeed, this will allow us to adapt the UIs to the users, and not the opposite. This should enhance the global acceptability and motivation of users facing a new interactive system.

3.4.2. A-posteriori user studies

In Potioc, we will conceive new hardware and software interfaces. To validate these UIs, and to improve them, we will conduct user experiments, as classically done in the field of HCI. This is a standard methodology that we currently follow (see Bibliography). We will do this in our lab, and in Cap Sciences.

Beyond the standard evaluation criteria that are based on performance for speed, accuracy, coordination, and so on, we will also consider other criteria that are more relevant for the Potioc project. Indeed, we will give a great importance to enjoyability, pleasure of use, accessibility, and so on. Consequently, we will need to redefine the standard way to evaluate UIs. Once again, our relationship with Cap Sciences will help us in such investigations. The use of questionnaires will be a way to better understand how an interface should be designed to reach a successful use. In addition, we will observe and analyze how visitors tend to interact with various interfaces we will propose. For example, we will collect information like the time spent on a given interactive system or the number of smiles recorded during an interaction process. The identification of good criteria to use for the evaluation of a popular 3D UI will be one of the research directions of our team.

Conducting such *a-posteriori* studies, in particular with experts of mediation, with new criteria of success, will be a second answer to our goal of evaluating the pleasure linked to the use of 3D UIs.

3.4.3. Real-time cognitive analysis

Classically, the user's subjective preferences for a given 3DUI are assessed using questionnaires. While these questionnaires provide important information, this is only a partial, biased, a-posteriori/a-priori measure, since they are collected before or after the 3D interaction process. When questionnaires are administered during 3D interaction, this interrupts and disturbs the user, hence biasing the evaluation. Moreover, while evaluating performance and usefulness is now well described and understood, evaluating the user's experiance and thus the system usability appears as much more difficult, with a lack of systematic and standard approaches. Ideally, we would like to measure the user response and subjective experience while he/she is using the 3DUI, i.e., in real-time and without interrupting him/her, in order to precisely identify the UI pros and cons. Questionnaires cannot provide such a measure.

Fortunately, it has been recently shown that BCI could be used in a passive way, to monitor the user's mental state. More precisely, recent results suggested that appropriately processed EEG signals could provide information about mental states such as error perception, attention or mental workload. As such, BCI are emerging as a new tool to monitor a user's mental state and brain responses to various stimuli, in real-time. In the Potioc project, we propose a completely new way to evaluate 3DUI: rather than relying only on questionnaires to estimate the user's subjective experience, we propose to exploit passive BCI to estimate the user's mental state in real-time, without interrupting nor disturbing him or her, while he/she is using the 3DUI. In particular, we aim at measuring and processing EEG and other biosignals (e.g., pulse, galvanic skin response, electromyogram) in real-time in order to estimate mental states such as interaction error potentials or workload/attention levels, among others. This will be used to finely identify how intuitive, easy-to-use and (ideally) enjoyable any given 3D UI is. More specifically, it will allow us to identify how, when and where the UI has flaws. Because the analysis will occur in real-time, we will potentially be able to modify the interface while the user is interacting. This should lead to a better understanding of 3D interaction. The work that will be achieved in this area could potentially also be useful for 2D interface design. However, since Potioc's main target is 3DUI, we will naturally focus the real-time cognitive evaluations on 3D contexts, with specific targets such as depth perception, or perception of 3D rotations.

This real-time cognitive analysis will be a third answer to reach the objectives of Potioc, which are to open 3D digital worlds to everyone by increasing the pleasure of use.

PRIMA Project-Team

3. Scientific Foundations

3.1. Context Aware Smart Spaces

Situation Models for Context Aware Systems and Services

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 [52], [53]. 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 [37], [36].

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 context (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 (more details at the end of this section).

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 (in a group of meeting participants, who is interacting with whom and thus which interaction groups are formed)

Currently situation models are constructed by hand. Our current 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 curently exploring reinforcement learning approaches to solve this problem.

With a reinforcement learning approach, the system is rewarded and punished by user reactions to system behaviors. A simplified stereotypic interaction model assures a initial behavior. 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 camera-man 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 (activity modeling, personal assistant). OMiSCID has facilitated cooperation by experts from within the PRIMA project as well as in projects with external partners.

Experiments with semantic service description and spontaneous service composition are conducted around the OMiSCID middleware. In these experiments, attention is paid to usability. A dedicated language has been designed to allow developers to describe the functionalities that their services provide. This language aims at simplifying existing semantic web services technologies to make them usable by a normal developer (i.e. that is not specialized in the semantic web). This language is named the User-oriented Functionality Composition Language (UFCL).

UFCL allows developers to specify three types of knowledge about services:

The knowledge that a service exposes a functionality like a “Timer” functionality for a service emitting message at a regular frequency.

The knowledge that a kind of functionality can be converted to another one. For example, a “Metronome” functionality issued from a music centered application can be seen as a “Timer” functionality.

The knowledge that a particular service is a factory and can instantiate other services on demand. A TimerFactory can for example start a new service with a “Timer” functionality with any desired frequency. Factories greatly helps in the deployment of service based applications. UFCL factories can also express the fact that they can compose existing functionalities to provide another one.

To bring the UFCL descriptions provided by the developers to life, a runtime has been designed to enable reasoning about what functionalities are available, what functionalities can be transformed to another one and what functionalities could be obtained by asking factories. The service looking for a particular functionality has just to express its need in term of functionalities and properties (e.g. a “Timer” with a frequency of 2Hz) and the runtime automates everything else: gathering of UFCL descriptions exposed by all running services, compilation of these descriptions to some rules in a rule-based system, reasoning and creation of a plan to obtained the desired functionality, and potentially invoking service factories to start the missing services.

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 [54]. 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 [56] 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 [62]. 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 [56], and normalized in orientation using direction of the first derivatives [62].

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" [57]. 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" [47] of Gaussian derivatives. The problem is to determine the appropriate orientation. In earlier work by PRIMA members Colin de Verdiere [34], Schiele [62] and Hall [50], proposed normalising the local jet independently at each pixel to the direction of the first derivatives calculated at the intrinsic scale. This has provided promising results for many view invariant image recognition tasks as described in the next section.

Color is a powerful discriminator for object recognition. Color images are commonly acquired in the Cartesian color space, RGB. The RGB color space has certain advantages for image acquisition, but is not the most appropriate space for recognizing objects or describing their shape. An alternative is to compute a Cartesian representation for chrominance, using differences of R, G and B. Such differences yield color opponent receptive fields resembling those found in biological visual systems.

Our work in this area uses a family of steerable color opponent filters developed by Daniela Hall [50]. These filters transform an (R,G,B), into a cartesian representation for luminance and chrominance (L,C1,C2). Chromatic Gaussian receptive fields are computed by applying the Gaussian derivatives independently to each of the three components, (L, C1, C2). The components C1 and C2 encodes the chromatic information in a Cartesian representation, while L is the luminance direction. Chromatic Gaussian receptive fields are computed by applying the Gaussian derivatives independently to each of the three components, (L, C1, C2). Permutations of RGB lead to different opponent color spaces. The choice of the most appropriate space depends on the chromatic composition of the scene. An example of a second order steerable chromatic basis is the set of color opponent filters shown in figure 1 .

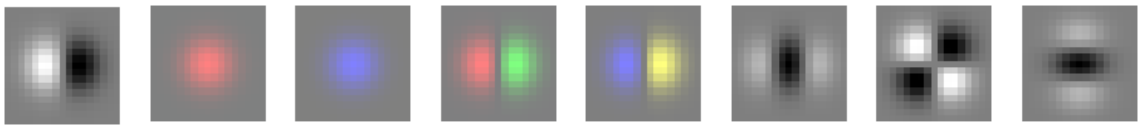


Figure 1. Chromatic Gaussian Receptive Fields ($G_x^L, G^{C_1}, G^{C_2}, G_x^{C_1}, G_x^{C_2}, G_{xx}^L, G_{xy}^L, G_{yy}^L$).

Key results in this area include

Fast, video rate, calculation of scale and orientation for image description with normalized chromatic receptive fields [37].

Real time indexing and recognition using a novel indexing tree to represent multi-dimensional receptive field histograms [60].

Robust visual features for face tracking [49], [48].

Affine invariant detection and tracking using natural interest lines [63].

Direct computation of time to collision over the entire visual field using rate of change of intrinsic scale [58].

We have achieved video rate calculation of scale and orientation normalised Gaussian receptive fields using an O(N) pyramid algorithm [37]. 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 at video rates. This method has been used in the MinImage project to provide real time detection, tracking, and identification of faces. It has also been used to provide techniques for estimating age and gender of people from their faces

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 will require 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 the 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 ICT 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 [45] has let a number of research groups to develop libraries of techniques for recognizing the elements of the FACS coding system [33]. 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 behaviors multimodal information. These are unconscious inmate cognitive processes that are vital to human communication and interaction. Recognition of social behaviors enables anticipation and improves the quality of interaction between humans. Among social behaviors, 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 [65]. 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 that, while such uni-modal methods may be suitable for static display [66] or robots in wide space area [55] they are not sufficient 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 behavior by the robot.

In our experiments, we use a kompai robot from Robosoft [32]. As an alternative to wearable physiological sensors (such as pulse bracelet Cardiocam, etc.) we inegrate multimodal features using a Kinect sensor (see figure 5). 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 [61].

Our multimodal approach has been confronted to a robot centered dataset for multimodal social signal processing recorded in a home-like environment [24]. 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 [61]: relative shoulder rotation, speed and facing visage are among crucial features for engagement detection.

3.5. End Users control over Smart Environment

Missing keywords.

Ubiquitous 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 functionalities that were not anticipated by system designers. A number of tools and techniques have been developed to support this view such as CAMP [64] or iCAP [44].

We adopt a 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 [38] than procedure in the sens of traditionnal programming languages.

Detecting and solving conflicts related to contradictory programms or goals.

REVES Project-Team

3. Scientific Foundations

3.1. 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. Plausible Rendering

3.1.1.1. 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" [37], 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 [27], but also simpler methods such as accessibility maps [34]. 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.

3.1.1.2. 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 [26], 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.

3.1.2. High Quality Rendering Using Simulation

3.1.2.1. 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 [35] which have been recently developed for sound research.

3.1.2.2. 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 [31], [30], [29] 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 [28]. Lazy evaluation, as well as hierarchical solutions, are clearly interesting avenues of research, although are probably quite application dependent.

3.1.2.3. 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.

3.1.2.4. 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 [32] [35]), 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 [33] 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).

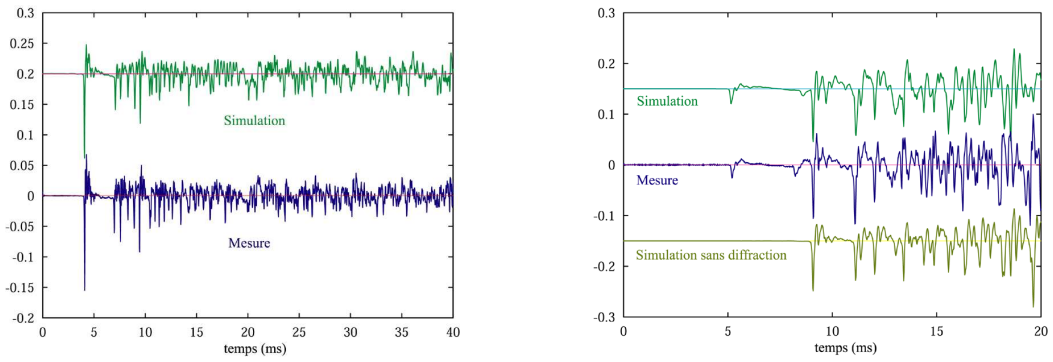


Figure 2. A comparison between a measurement (left) of the sound pressure in a given location of the "Bell Labs Box", a simple test environment built at Bell Laboratories, and a high-quality simulation based on a beam-tracing engine (right). Simulations include effects of reflections off the walls and 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 [36].

SEMAGRAMME Team

3. Scientific Foundations

3.1. Fondation

The present proposal 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

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

(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, [28] 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. Scientific Foundations

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 the 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 [3]. 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 principal component analysis (PCA) and independent component analysis (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 unary 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 the 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 [4]. This detection is also influenced by prior knowledge about the environment of the scene[5]. Implicit assumptions related to Prior knowledge or beliefs form 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[6]. 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:

$$S(\overrightarrow{x}) = \frac{1}{p(\overrightarrow{v}_l \,|\, \overrightarrow{v}_c)} \times p(s, \overrightarrow{x} \,|\, \overrightarrow{v}_c) \tag{114}$$

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.

[3] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization", Nature 401, 6755, (Oct. 1999), pp. 788-791.

[4] L. Itti and C. Koch, "Computational Modelling of Visual Attention" , Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203, 2001.

[5] J. Henderson, "Regarding scenes", Directions in Psychological Science, vol. 16, pp. 219-222, 2007.

[6] 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 [7] 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[8] (SW) and Wyner-Ziv [9] (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.

[7] T. M. Cover and J. A. Thomas, Elements of Information Theory, Second Edition, July 2006.

[8] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources." IEEE Transactions on Information Theory, 19(4), pp. 471-480, July 1973.

[9] 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. Scientific Foundations

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 [39] and Hippocratic databases [30]. In the last years, many other policy languages have been proposed for different application scenarios, including EPAL [44], XACML [41] and WSPL [34]. 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 behaviour 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) [36]. 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 (R. Agrawal : IBM Almaden, C.C. Aggarwal: IBM Watson), 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 [40]. 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 (R.N.Wright : Stevens Institute of Tech. / Rutgers Univ, NJ, V. Shmatikov : Univ Austin Texas), on Secure Multi-party Computing for Privacy Preserving Data Mining (J. Vaidya : Rutgers Univ, C. Clifton : Purdue Univ) and on distributed PPDP algorithms (D. DeWitt : Univ Wisconsin, K. Lefevre : Univ Michigan, J. Vaidya : Rutgers Univ, C. Clifton : Purdue Univ) while none of them share the same architectural hypothesis as us.

STARS Team

3. Scientific Foundations

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 architecture 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 sensors (e.g. Kinect) 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 framework1pt 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 analysing 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 two following research axes: high level understanding (to recognize the activities of physical objects based on high level activity models) and learning (how to learn the models needed for activity recognition).

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. We have also 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

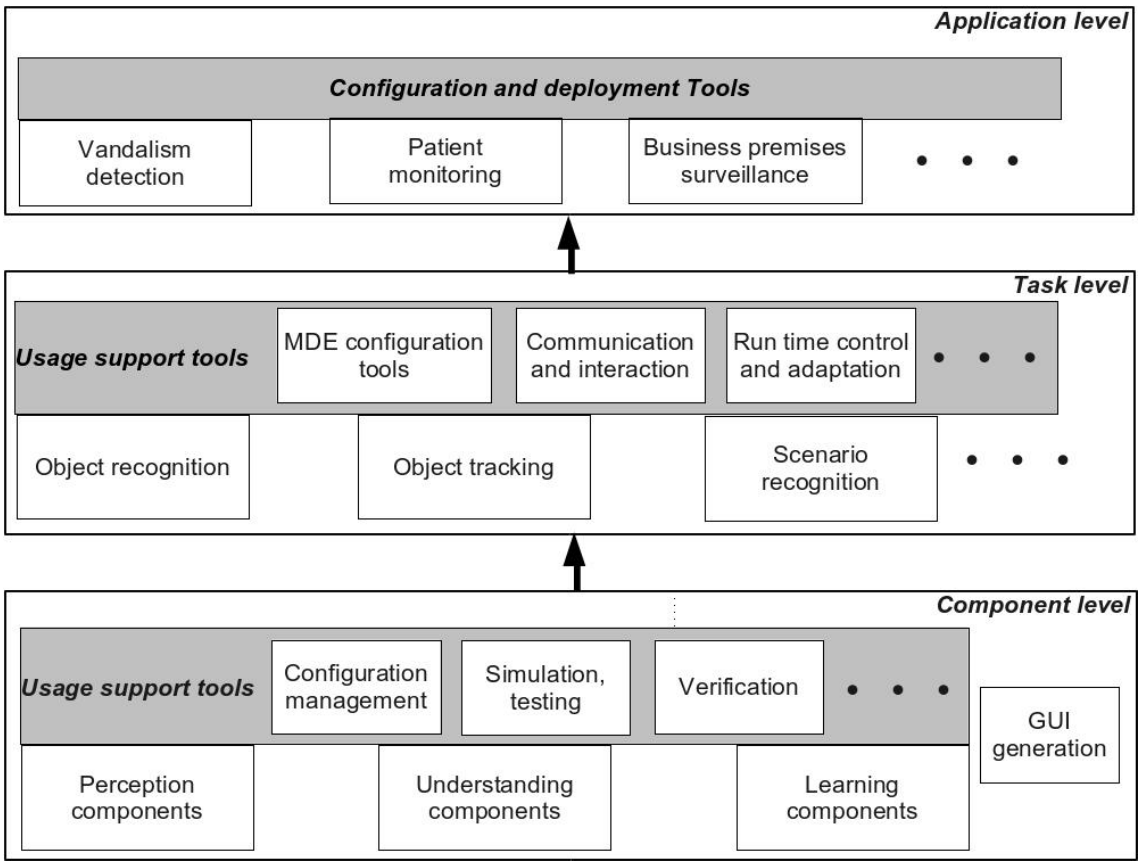


Figure 4. Global Architecture of an Activity Recognition The grey areas contain software engineering support modules whereas the other modules correspond to software components (at Task and Component levels) or to generated systems (at Application level).

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 [15], 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 or a cognitive 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 parametrized 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, 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 [54]. 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 [70]. 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 [51], [80], [72].

TEXMEX Project-Team

3. Scientific Foundations

3.1. Image description

In most contexts where images are to be compared, a direct comparison is impossible. Images are compressed in different formats, most formats are error-prone, images are re-sized, cropped, etc. The solution consists in computing descriptors, which are invariant to these transformations.

The first description methods associate a unique global descriptor with each image, *e.g.*, a color histogram or correlogram, a texture descriptor. Such descriptors are easy to compute and use, but they usually fail to handle cropping and cannot be used for object recognition. The most successful approach to address a large class of transformations relies on the use of local descriptors, extracted on regions of interest detected by a detector, for instance the Harris detector [82] or the Difference of Gaussian method proposed by David Lowe [84].

The detectors select a square, circular or elliptic region that is described in turn by a patch descriptor, usually referred to as a local descriptor. The most established description method, namely the SIFT descriptor [84], was shown robust to geometric and photometric transforms. Each local SIFT descriptor captures the information provided by the gradient directions and intensities in the region of interest in each region of a 4×4 grid, thereby taking into account the spatial organization of the gradient in a region. As a matter of fact, the SIFT descriptor has became a standard for image and video description.

Local descriptors can be used in many applications: image comparison for object recognition, image copy detection, detection of repeats in television streams, etc. While they are very reliable, local descriptors are not without problems. As many descriptors can be computed for a single image, a collection of one million images generates in the order of a billion descriptors. That is why specific indexing techniques are required. The problem of taking full advantage of these strong descriptors on a large scale is still an open and active problem. Most of the recent techniques consists in computing a global descriptor from local ones, such as proposed in the so-called bag-of-visual-word approach [89]. Recently, global description computed from local descriptors has been shown successful in breaking the complexity problem. We are active in designing methods that aggregate local descriptors into a single vector representation without loosing too much of the discriminative power of the descriptors.

3.2. Corpus-based text description and machine learning

Our work on textual material (textual documents, transcriptions of speech documents, captions in images or videos, etc.) is characterized by a chiefly corpus-based approach, as opposed to an introspective one. A corpus is for us a huge collection of textual documents, gathered or used for a precise objective. We thus exploit specialized (abstracts of biomedical articles, computer science texts, etc.) or non specialized (newspapers, broadcast news, etc.) collections for our various studies. In TEXMEX, according to our applications, different kinds of knowledge can be extracted from the textual material. For example, we automatically extract terms characteristic of each successive topic in a corpus with no a priori knowledge; we produce representations for documents in an indexing perspective [88]; we acquire lexical resources from the collections (morphological families, semantic relations, translation equivalences, etc.) in order to better grasp relations between segments of texts in which a same idea is expressed with different terms or in different languages...

In the domain of the corpus-based text processing, many researches have been undergone in the last decade. While most of them are essentially based on statistical methods, symbolic approaches also present a growing interest [78]. For our various problems involving language processing, we use both approaches, making the most of existing machine learning techniques or proposing new ones. Relying on advantages of both methods, we aim at developing machine learning solutions that are automatic and generic enough to make it possible to extract, from a corpus, the kind of elements required by a given task.

3.3. Stochastic models for multimodal analysis

Describing multimedia documents, *i.e.*, documents that contain several modalities (*e.g.*, text, images, sound) requires taking into account all modalities, since they contain complementary pieces of information. The problem is that the various modalities are only weakly synchronized, they do not have the same rate and combining the information that can be extracted from them is not obvious. Of course, we would like to find generic ways to combine these pieces of information. Stochastic models appear as a well-dedicated tool for such combinations, especially for image and sound information.

Markov models are composed of a set of states, of transition probabilities between these states and of emission probabilities that provide the probability to emit a given symbol at a given state. Such models allow generating sequences. Starting from an initial state, they iteratively emit a symbol and then switch in a subsequent state according to the respective probability distributions. These models can be used in an indirect way. Given a sequence of symbols (called observations), hidden Markov models (HMMs, [87]) aim at finding the best sequence of states that can explain this sequence. The Viterbi algorithm provides an optimal solution to this problem.

For such HMMs, the structure and probability distributions need to be a priori determined. They can be fixed manually (this is the case for the structure: number of states and their topology), or estimated from example data (this is often the case for the probability distributions). Given a document, such an HMM can be used to retrieve its structure from the features that can be extracted. As a matter of fact, these models allow an audiovisual analysis of the videos, the symbols being composed of a video and an audio component.

Two of the main drawbacks of the HMMs is that they can only emit a unique symbol per state, and that they imply that the duration in a given state follows an exponential distribution. Such drawbacks can be circumvented by segment models [86]. These models are an extension of HMMs were each state can emit several symbols and contains a duration model that governs the number of symbols emitted (or observed) for this state. Such a scheme allows us to process features at different rates.

Bayesian networks are an even more general model family. Static Bayesian networks [80] are composed of a set of random variables linked by edges indicating their conditional dependency. Such models allow us to learn from example data the distributions and links between the variables. A key point is that both the network structure and the distributions of the variables can be learned. As such, these networks are difficult to use in the case of temporal phenomena. Dynamic Bayesian [85] networks are a generalization of the previous models. Such networks are composed of an elementary network that is replicated at each time stamp. Duration variable can be added in order to provide some flexibility on the time processing, like it was the case with segment models. While HMMs and segment models are well suited for dense segmentation of video streams, Bayesian networks offer better capabilities for sparse event detection. Defining a trash state that corresponds to non event segments is a well known problem is speech recognition: computing the observation probabilities in such a state is very difficult.

3.4. Multidimensional indexing techniques

Techniques for indexing multimedia data are needed to preserve the efficiency of search processes as soon as the data to search in becomes large in volume and/or in dimension. These techniques aim at reducing the number of I/Os and CPU cycles needed to perform a search. Multi-dimensional indexing methods either perform exact nearest neighbor (NN) searches or approximate NN-search schemes. Often, approximate techniques are faster as speed is traded off against accuracy.

Traditional multidimensional indexing techniques typically group high dimensional features vectors into cells. At querying time, few such cells are selected for searching, which, in turn, provides performance as each cell contains a limited number of vectors [79]. Cell construction strategies can be classified in two broad categories: *data-partitioning* indexing methods that divide the data space according to the distribution of data, and *space-partitioning* indexing methods that divide the data space along predefined lines and store each descriptor in the appropriate cell.

Unfortunately, the “curse of dimensionality” problem strongly impacts the performance of many techniques. Some approaches address this problem by simply relying on dimensionality reduction techniques. Other approaches abort the search process early, after having accessed an arbitrary and predetermined number of cells. Some other approaches improve their performance by considering approximations of cells (with respect to their true geometry for example).

Recently, several approaches make use of quantization operations. This, somehow, transforms costly nearest neighbor searches in multidimensional space into efficient uni-dimensional accesses. One seminal approach, the LSH technique [81], uses a structured scalar quantizer made of projections on segmented random lines, acting as spatial locality sensitive hash-functions. In this approach, several hash functions are used such that co-located vectors are likely to collide in buckets. Other approaches use unstructured quantization schemes, sometimes together with a vector aggregation mechanism [89] to boost performance.

3.5. Data mining methods

Data Mining (DM) is the core of knowledge discovery in databases whatever the contents of the databases are. Here, we focus on some aspects of DM we use to describe documents and to retrieve information. There are two major goals to DM: description and prediction. The descriptive part includes unsupervised and visualization aspects while prediction is often referred to as supervised mining.

The description step very often includes feature extraction and dimensional reduction. As we deal mainly with contingency tables crossing "documents and words", we intensively use factorial correspondence analysis. "Documents" in this context can be a text as well as an image.

Correspondence analysis is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information, which is similar in nature to those produced by factor analysis techniques, and they allow one to explore the structure of categorical variables included in the table. The most common kind of table of this type is the two-way frequency cross-tabulation table. There are several parallels in interpretation between correspondence analysis and factor analysis: suppose one could find a lower-dimensional space, in which to position the row points in a manner that retains all, or almost all, of the information about the differences between the rows. One could then present all information about the similarities between the rows in a simple 1, 2, or 3-dimensional graph. The presentation and interpretation of very large tables could greatly benefit from the simplification that can be achieved via correspondence analysis (CA).

One of the most important concepts in CA is inertia, *i.e.*, the dispersion of either row points or column points around their gravity center. The inertia is linked to the total Pearson χ^2 for the two-way table. Some rows and/or some columns will be more important due to their quality in a reduced dimensional space and their relative inertia. The quality of a point represents the proportion of the contribution of that point to the overall inertia that can be accounted for by the chosen number of dimensions. However, it does not indicate whether or not, and to what extent, the respective point does in fact contribute to the overall inertia (χ^2 value). The relative inertia represents the proportion of the total inertia accounted for by the respective point, and it is independent of the number of dimensions chosen by the user. We use the relative inertia and quality of points to characterize clusters of documents. The outputs of CA are generally very large. At this step, we use different visualization methods to focus on the most important results of the analysis.

In the supervised classification task, a lot of algorithms can be used; the most popular ones are the decision trees and more recently the Support Vector Machines (SVM). SVMs provide very good results in supervised classification but they are used as "black boxes" (their results are difficult to explain). We use graphical methods to help the user understanding the SVM results, based on the data distribution according to the distance to the separating boundary computed by the SVM and another visualization method (like scatter matrices or parallel coordinates) to try to explain this boundary. Other drawbacks of SVM algorithms are their computational cost and large memory requirement to deal with very large datasets. We have developed a set of incremental and parallel SVM algorithms to classify very large datasets on standard computers.

VR4I Team

3. Scientific Foundations

3.1. Panorama

Our main concern is to allow real users to interact naturally within shared virtual environments as interaction can be the result of an individual interaction of one user with one object or a common interaction of several users on the same object. The long-term purpose of the project is to propose interaction modalities within virtual environments that bring **acting in Virtual Reality as natural as acting in reality**.

Complex physically based models have to be proposed to represent the virtual environment, complex multi-modal interaction models have to be proposed to represent natural activity and complex collaborative environments have to be proposed to ensure effective collaborative interactions.

The long term objectives of VR4i are:

- Improving the accuracy of the virtual environment representation for more interactivity and better perception of the environment;
- Improving the multi-modal interaction for more natural interactions and better perception of the activity;
- Improving the use of virtual environments for real activity and open to human science for evaluation and to engineering science for applications.

Thus, we propose three complementary research axes:

- Physical modeling and simulation of the environment
- Multimodal immersive interaction
- Collaborative work in Collaborative Virtual Environments (CVE)

3.2. Physical modeling and simulation

The first aspect is the modeling and the simulation of the virtual world that represents properly the physical behavior of the virtual world that sustains a natural interaction through the different devices. The main challenge is the search of the trade-off between accuracy and performance to allow effective manipulation, in interactive time, by the user. This trade-off is a key point while the user closes the interaction loop. Namely, the accuracy of the simulation drives the quality of the phenomenon to perceive and the performance drives the sensori-motor feelings of the user. Proposing new controlled algorithms for physical based simulation of the virtual world is certainly a key point for meeting this trade-off. We believe that the mechanical behavior of objects as to be more studied and to be as close as possible to their real behavior. The devices may act as a both way filter on the action and on the perception of the simulated world, but improving the representation of rigid objects submitted to contact, of deformable objects, of changing state object and of environments that include mixed rigid and deformable objects is needed in order to compute forces and positions that have a physical meaning. The interaction between tools and deformable objects is still a challenge in assembly applications and in medical applications. The activity of the user in interaction with the immersive environment will allow to provide method to qualify the quality of the environment and of the interaction by proposing a bio-mechanical user's Alter Ego. We believe that the analysis of the forces involved during an immersive activity will give us keys to design more acceptable environments. As the goal is to achieve more and more accurate simulation that will require more and more computation time, the coupling between physical modeling and related simulation algorithms is of first importance. Looking for genericity will ensure correct deployment on new advanced hardware platforms that we will use to ensure adapted performance. The main aim of this topic is to improve the simulation accuracy satisfying the simulation time constraints for improving the naturalness of interactions.

3.3. Multimodal immersive interaction

The second aspect concerns the design and evaluation of novel approaches for multimodal immersive interaction with virtual environments.

We aim at improving capabilities of selection and manipulation of virtual objects, as well as navigation in the virtual scene and control of the virtual application. We target a wide spectrum of sensory modalities and interfaces such as tangible devices, haptic interfaces (force-feedback, tactile feedback), visual interfaces (e.g., gaze tracking), locomotion and walking interfaces, and brain-computer interfaces. We consider this field as a strong scientific and technological challenge involving advanced user interfaces, but also as strongly related to user's perceptual experience. We promote a perception-based approach for multimodal interaction, based on collaborations with laboratories of the Perception and Neuroscience research community.

The introduction of a third dimension when interacting with a virtual environment makes inappropriate most of the classical techniques used successfully in the field of 2D interaction with desktop computers up to now. Thus, it becomes successfully used to design and evaluate new paradigms specifically oriented towards interaction within 3D virtual environments.

We aim at improving the immersion of VR users by offering them natural ways for navigation, interaction and application control, as these are the three main tasks within 3D virtual environments. Here we consider interactions as multimodal interactions, as described in the previous section. We also want to make the users forget their physical environment in benefit of the virtual environment that surrounds them and contribute to improve the feeling of immersion and of presence. To achieve this goal, we must ensure that users can avoid collisions with their surrounding real environment (the screens of the rendering system, the walls of the room) and can avoid lost of interaction tracking (keeping the user within the range of the physical interaction devices). To do that, we propose to take into account the surrounding real physical environment of the user and to include it in the virtual environment through a virtual representation. This explicit model of the real environment of the users will help users to forget it: throughout this model, the user will be aware (with visual, auditive or haptic feedback) of these virtual objects when he comes near their boundaries. We also have to investigate which physical limitations are the most important ones to perceive, and what are the best ways to make the users aware of their physical limitations.

3.4. Collaborative work in CVE's

The third aspect is to propose Collaborative Virtual Environments for several local or distant users. In these environments, distant experts could share their expertise for project review, for collaborative design or for analysis of data resulting from scientific computations in HPC context. Sharing the virtual environment is certainly a key point that leads to propose new software architectures ensuring the data distribution and the synchronization of the users.

In terms of interaction, new multi-modal interaction metaphors have to be proposed to tackle with the awareness of other users? activity. Here it is important to see a virtual representation of the other users, of their activity, and of the range of their action field, in order to better understand both their potential and their limitation for collaboration: what they can see, what they can reach, what their interaction tools are and which possibilities they offer.

Simultaneous collaborative interactions upon the same data through local representations of these data should be tackled by new generic algorithms dedicated to consistency management. Some solutions have to be proposed for distant collaboration, where it is not possible any more to share tangible devices to synchronize co-manipulation: we should offer some new haptic rendering to enforce users' coordination. Using physics engines for realistic interaction with virtual objects is also a challenge if we want to offer low latency feedback to the users. Indeed, the classical centralized approach for physics engines is not able to offer fast feedback to distant users, so this approach must be improved.

WAM Project-Team

3. Scientific Foundations

3.1. XML Processing

Participants: Melisachew Chekol, Pierre Genevès, Nils Gesbert, Nicola Guido, Muhammad Junedi, Nabil Layaïda, Manh-Toan Nguyen, Vincent Quint.

Extensible Markup Language (XML) has gained considerable interest from industry, and plays now a central role in modern information system infrastructures. In particular, XML is the key technology for describing, storing, and exchanging a wide variety of data on the web. The essence of XML consists in organizing information in tree-tagged structures conforming to some constraints which are expressed using type languages such as DTDs, XML Schemas, and Relax NG.

There still exist important obstacles in XML programming, especially in the areas of performance and reliability. Programmers are given two options: domain-specific languages such as XSLT, or general-purpose languages augmented with XML application programming interfaces such as the Document Object Model (DOM). Neither of these options is a satisfactory answer to performance and reliability issues, nor is there even a trade-off between the two. As a consequence, new paradigms are being proposed which all have the aim of incorporating XML data as first-class constructs in programming languages. The hope is to build a new generation of tools that are capable of taking reliability and performance into account at compile time.

One of the major challenges in this line of research is to develop automated and tractable techniques for ensuring static type safety and optimization of programs. To this end, there is a need to solve some basic reasoning tasks that involve very complex constructions such as XML types (regular tree types) and powerful navigational primitives (XPath expressions or CSS selectors). In particular, every future compiler of XML programs will have to routinely solve problems such as:

- XPath query emptiness in the presence of a schema: if one can decide at compile time that a query is not satisfiable, then subsequent bound computations can be avoided
- query equivalence, which is important for query reformulation and optimization
- path type-checking, for ensuring at compile time that invalid documents can never arise as the output of XML processing code.

All these problems are known to be computationally heavy (when decidable), and the related algorithms are often tricky.

We have developed an XML/XPath static analyzer based on a new logic of finite trees. This analyzer consists of:

- compilers that allow XML types, XPath queries, and CSS selectors to be translated into this logic
- an optimized logical solver for testing satisfiability of a formula of this logic.

The benefit of these compilers is that they allow one to reduce all the problems listed above, and many others too, to logical satisfiability. This approach has a couple of important practical advantages. First of all, one can use the satisfiability algorithm to solve all of these problems. More importantly, one could easily explore new variants of these problems, generated for example by the presence of different kinds of type or schema information, with no need to devise a new algorithm for each variant.

3.2. Multimedia Models and Languages

Participants: Nicolas Hairon, Yohan Lasorsa, Nabil Layaïda, Jacques Lemordant, Vincent Quint, Cécile Roisin.

We have participated in the international endeavor for defining a standard multimedia document format for the web that accommodates the constraints of different types of terminals. SMIL is the main outcome of this work. It focuses on a modular and scalable XML format that combines efficiently the different dimensions of a multimedia web document: synchronization, layout and linking. Our current work on multimedia formats follows the same trend.

With the advent of HTML5 and its support in all popular browsers, HTML is becoming an important multimedia language. Video and audio can now be embedded in HTML pages without worrying about the availability of plugins. However, animation and synchronization of a HTML5 page still require programming skills. To address this issue, we are developing a scheduler that allows HTML documents to be animated and synchronized in a purely declarative way. This work is based on the SMIL Timing and Synchronization module and the SMIL Timesheets specification. The scheduler is implemented in JavaScript, which makes it usable in any browser. Timesheets can also be used with other XML document languages, such as SVG for instance.

Audio is the poor relation in the web format family. Most contents on the web may be represented in a structured way, such as text in HTML or XML, graphics in SVG, or mathematics in MathML, but sound was left aside with low-level representations that basically only encode the audio signal. Our work on audio formats aims at allowing sound to be on a par with other contents, in such a way it could be easily combined with them in rich multimedia documents that can then be processed safely in advanced applications. More specifically, we have participated in IAsig (Interactive Audio special interest group), an international initiative for creating a new format for interactive audio called iXMF (Interactive eXtensible Music Format). We are now developing A2ML, an XML format for embedded interactive audio, deriving from well-established formats such as iXMF and SMIL. We use it in augmented environments (see section 3.4), where virtual, interactive, 3D sounds are combined with the real sonic environment.

Regarding discrete media objects in multimedia documents, popular document languages such as HTML can represent a very broad range of documents, because they contain very general elements that can be used in many different situations. This advantage comes at the price of a low level of semantics attached to the structure. The concepts of microformats and semantic HTML were proposed to tackle this weakness. More recently, RDFa and microdata were introduced with the same goal. These formats add semantics to web pages while taking advantage of the existing HTML infrastructure. With this approach new applications can be deployed smoothly on the web, but authors of web pages have very little help for creating and encoding this kind of semantic markup. A language that addresses these issues is developed and implemented in WAM. Called XTiger, its role is to specify semantically rich XML languages in terms of other, less expressive XML languages, such as HTML. Recent extensions to the language make it now usable also to edit pure XML documents and to define their structure model (see section 3.3).

3.3. Multimedia Authoring

Participants: Nicolas Hairon, Yohan Lasorsa, Jacques Lemordant, David Liodenot, Vincent Quint, Mathieu Razafimahazo, Cécile Roisin.

3.3.1. Structured editing

Multimedia documents are considered through several kinds of structures: logical organization, layout, time, linking, animations. We are working on techniques that allow authors of such documents to manipulate all these structures in homogeneous environments. The main objective is to support new advances in document formats without making the authoring task more complex. The key idea is to present simultaneously several views of the document, each view putting the emphasis on a particular structure, and to allow authors to manipulate each view directly and efficiently. As the various structures of a document are not independent from each other, views are “synchronized” to reflect in each of them the consequences of every change made in a particular view. The XML markup, although it can be accessed at any time, is handled by the tools, and authors do not have to worry about syntactical issues.

3.3.2. Template-driven editing

We have more recently experimented another way to edit highly structured XML documents without the usual complexity of the most common XML editors. The novelty of the approach is to use templates instead of XML schemas or DTDs, and to run the editor as a web application, within the browser. This way, it is much easier to create new document types and to provide an editing environment for these document types, that any web user can instantly use. This lightweight approach to XML editing complements the previous approach by covering new categories of XML applications.

3.4. Augmented Environments

Participants: Yohan Lasorsa, Jacques Lemordant, David Liodenot, Thibaud Michel, Mathieu Razafimahazo.

The term Augmented Environments refers collectively to ubiquitous computing, context-aware computing, and intelligent environments. The goal of our research on these environments is to introduce personal Augmented Reality (AR) devices, taking advantage of their embedded sensors. We believe that personal AR devices such as mobile phones or tablets will play a central role in augmented environments. These environments offer the possibility of using ubiquitous computation, communication, and sensing to present context-sensitive information and services to the user.

AR applications often rely on 3D content and employ specialized hardware and computer vision techniques for both tracking and scene reconstruction. Our approach tries to seek a balance between these traditional AR contexts and what has come to be known as mobile AR browsing. It first acknowledges that mobile augmented environment browsing does not require that 3D content be the primary means of authoring. It provides instead a method for HTML5 and audio content to be authored, positioned in the surrounding environments and manipulated as freely as in modern web browsers.

Many service providers of augmented environments desire to create innovative services. Accessibility of buildings is one example we are involved in. However, service providers often have to strongly rely on experience, intuition, and tacit knowledge due to lack of tools on which to base a scientific approach. Augmented environments offer the required rigorous approach that enables Evidence-Based Services (EBS) if adequate tools for AR technologies are designed. Service cooperation through exchange of normalized real-time data or data logs is one of these tools, together with sensor data streams fusion inside an AR mobile browser. EBS can improve the performance of real-world sensing, and conversely EBS models authoring and service operation can be facilitated by real-world sensing.

The applications we use to elaborate and validate our concepts are pedestrian navigation for visually impaired people and applications for cultural heritage visits. On the authoring side, we are interested in interactive indoor modeling, audio mobile mixing, and formats for Points of Interest. Augmented environment services we consider are, among others, behavior analysis for accessibility, location services, and indoor geographical information services.

WILLOW Project-Team

3. Scientific Foundations

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 [1] 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.

for free for academics, and licensing negotiations with several companies are under way.

Our recent work (Russel *et al.*, 2011) has 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. This direction is currently being continued in the PhD work of Mathieu Aubry. Our other current work outlined in detail in Section 6.1 is focused on (i) recovering indoor scene geometry from observations of person-object interactions video, (ii) visual place recognition in structured databases, where images are geotagged and organized in a graph, and (iii) developing a discriminative clustering approach able to discover geographically representative image elements from Google Street View imagery using only weak geographic supervision.

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 6.2), focuses on the two problems described next.

[1] The patent "Match, Expand, and Filter Technique for Multi-View Stereopsis," was issued December 11, 2012 and assigned patent number 8,331,615.

3.2.1. Learning image and object models.

Learning sparse representations of images has been the topic of much recent research. It has been used for instance for image restoration (e.g., Mairal *et al.*, 2007) and it has been generalized to discriminative image understanding tasks such as texture segmentation, category-level edge selection and image classification (Mairal *et al.*, 2008). We have also developed fast and scalable optimization methods for learning the sparse image representations, and developed a software called SPAMS (SPArse Modelling Software) presented in Section 5.2 . The work of J. Mairal is summarized in his thesis (Mairal, 2010). The most recent work has focused on developing a general formulation for supervised dictionary learning and investigating methods to learn better mid-level features for recognition.

3.2.2. Category-level object/scene recognition and segmentation

Another significant strand of our research has focused on the extremely challenging goals of category-level object/scene recognition and segmentation. Towards these goals, we have developed: (i) strongly-supervised deformable part-based model for object recognition and localization, (ii) a MRF model for segmentation of text in natural scenes, and (iii) algorithms for multi-class cosegmentation using a novel energy-minimization approach based on the developed convex relaxation for weakly supervised classifiers.

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.

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 6.3 , has focused on (i) developing a geometrical model for removing image blur due to camera shake, (ii) preparing an online image deblurring demo, and (iii) developing new formulation for image deblurring cast as a multi-label energy minimization 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 6.4 .

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 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.

3.4.4. Modeling and recognizing person-object and person-scene interactions.

Actions of people are tightly coupled with their environments and surrounding objects. Moreover, object function can be learned and recognized from observations of person-object interactions in video and still images. Designing and learning models for person-object interactions, however, is a challenging task due to both (i) the huge variability in visual appearance and (ii) the lack of corresponding annotations. We address this problem by developing weakly-supervised techniques enabling learning interaction models from long-term observations of people in natural indoor video scenes such as obtained from time-lapse videos on YouTube. We also explore stereoscopic information in 3D movies to learn better models for people in video including person detection, segmentation, pose estimation, tracking and action recognition.

WIMMICS Team

3. Scientific Foundations

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 the 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 the 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 the 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 for software to exploit them in their processing. 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. Scientific Foundations

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, 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) [13]. 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 [9] 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. 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 wrt. 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.

Current cloud data management (NOSQL) solutions 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 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 these NOSQL technologies to fit our needs for scientific workflow management and scalable data analysis.

3.4. 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 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.

3.5. Metadata Integration

Nowdays, 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 identifying 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 term 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, so does their diversity and heterogeneity. As a result, schema/ontology matching has become a prominent and challenging topic [4].

3.6. 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 [12].** 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 recently was 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 time response 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 maintaining 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.

3.7. 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 retrieve efficiently 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 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.
- **P2P content-based retrieval**. Content-based P2P retrieval methods appeared recently as a promising solution to manage masses of data distributed over large social 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 P2P content-based retrieval and mining methods such as k-nn graphs construction, large-scale supervised retrieval or multi-source clustering.