RESEARCH CENTER
Grenoble - Rhône-Alpes

FIELD

Activity Report 2012

Section Scientific Foundations

Edition: 2013-04-24

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.

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?

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.

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.

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

$$\begin{cases} \dot{x} = f(x,u,\lambda)\,, \\ 0 \le y \perp \lambda \ge 0\,, \\ g(y,\lambda,x,u,t) = 0\,, \\ \text{re-initialization law of the state } x(\cdot)\ , \end{cases} \tag{1}$$

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].

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,

$$f(y_i \mid \theta) = \sum_{k=1}^{K} P(z_i = k \mid \theta) f(y_i \mid z_i, \theta) \ . \tag{2}$$

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).

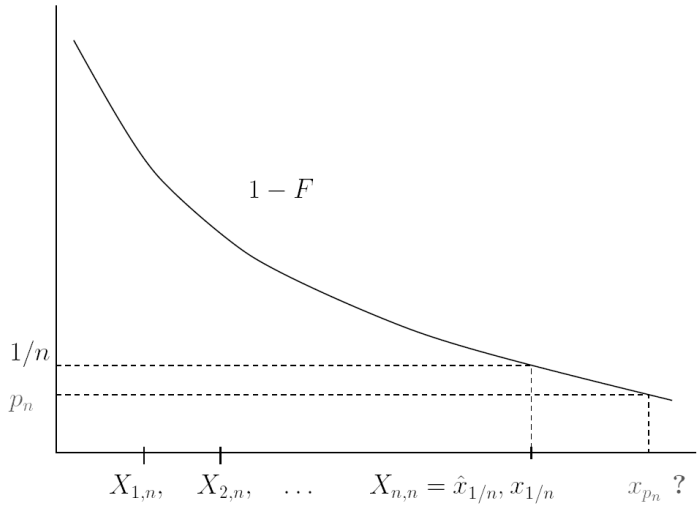


Figure 1. The curve represents the survival function $x \to P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\widehat{x}_{1/n} = X_{n,n}$. As illustrated in the figure, to estimate p_n-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.

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:

$$P(X > x) = x^{-1/\theta}\ell(x),\ x > x_0 > 0, \tag{3}$$

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

$$\frac{\ell(tx)}{\ell x} \to 1 \text{ as } x \to \infty \tag{4}$$

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:

$$P(X > x) = \exp\{-x^{\theta}\ell(x)\},\ x > x_0 > 0. \tag{5}$$

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].

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.

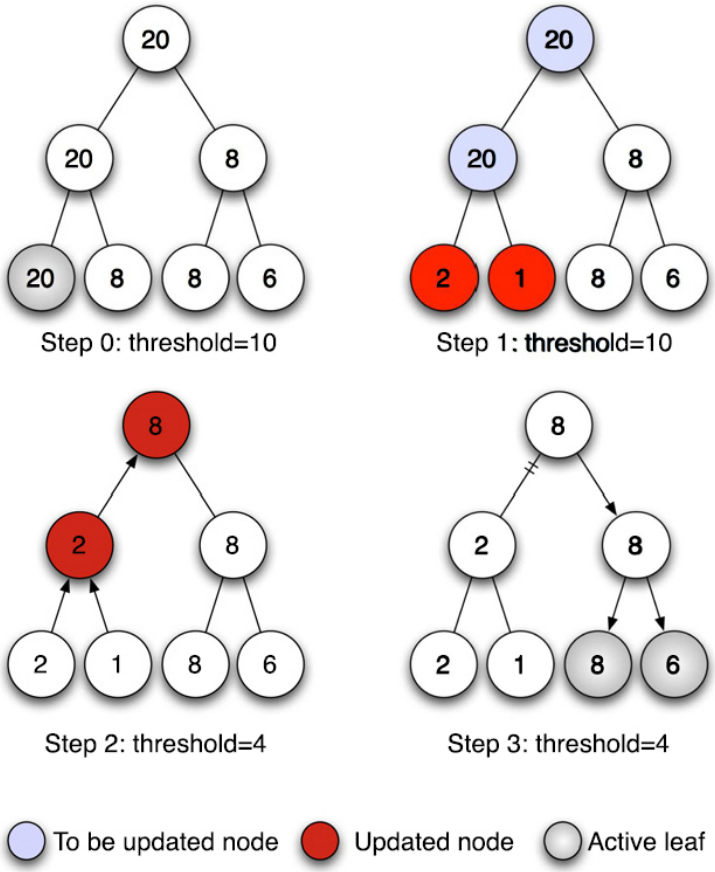


Figure 2. Adaptive Cartesian mechanics.

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.

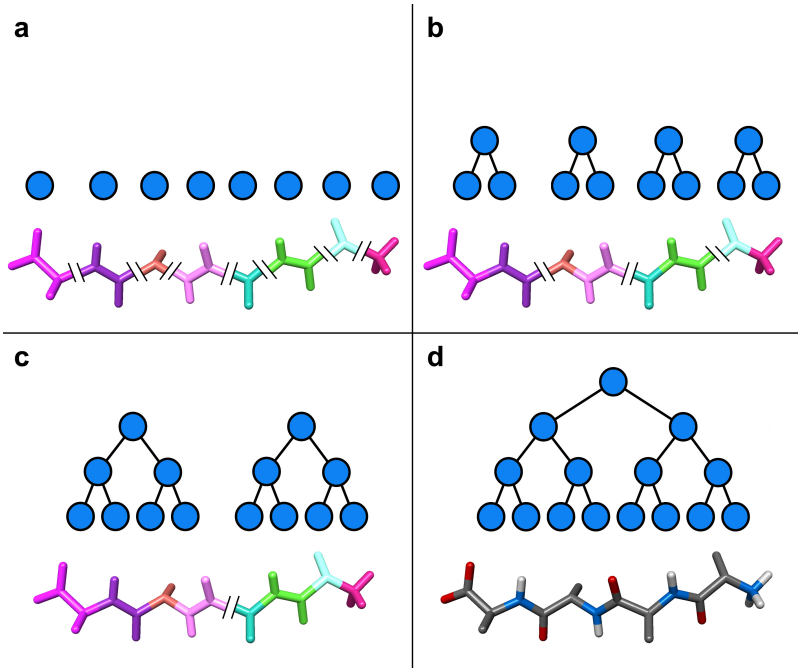


Figure 3. The assembly tree associated to a short polyalanin.

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:

$$\mathcal{A}(C) = (\mathbf{f}^C)^T \Psi^C \mathbf{f}^C + (\mathbf{f}^C)^T \mathbf{p}^C + \eta^C, \tag{6}$$

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

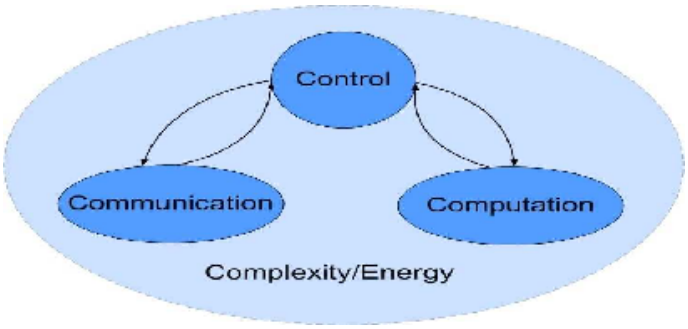


Figure 2. Relation of the NCS area with the fields of: Control, Communication, Computation.

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.

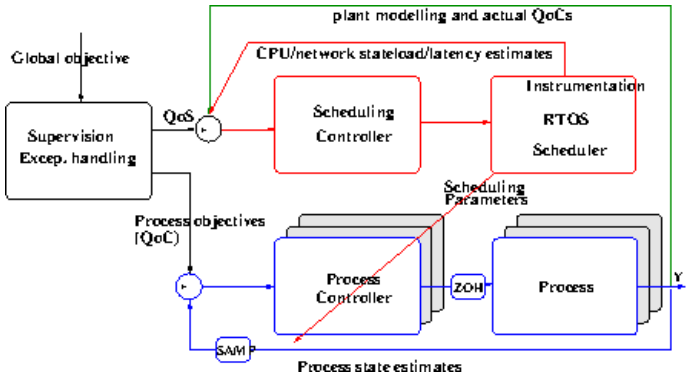


Figure 3. Hierarchical control structure.

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).

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.

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.

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.

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).

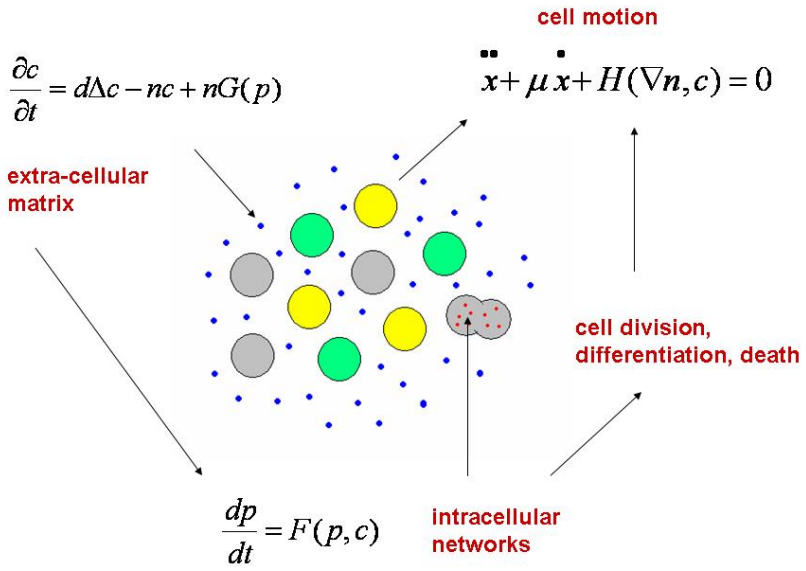


Figure 1. Schema of multi-scale models of cell dynamics: DPD-PDE-ODE models.

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.

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.

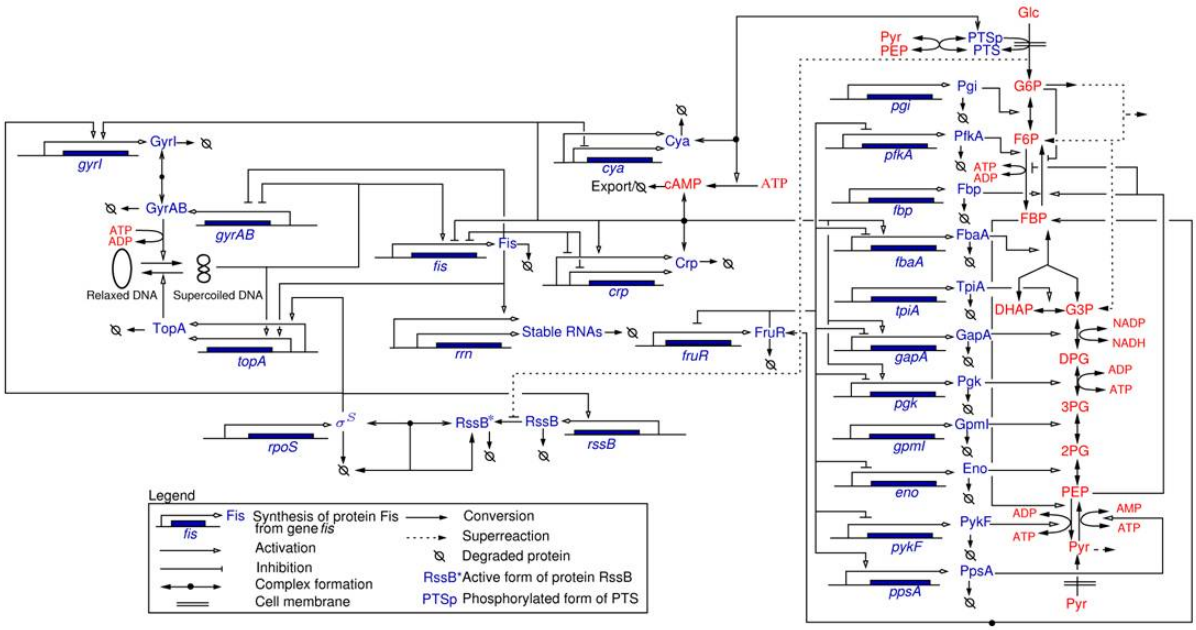


Figure 2. Network of key genes, proteins, and regulatory interactions involved in the carbon assimilation network in E. coli (Baldazzi et al., PLoS Computational Biology, 6(6):e1000812, 2010). The metabolic part includes the glycolysis/gluconeogenesis pathways as well as a simplified description of the 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.

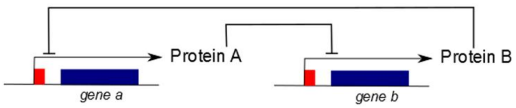
3.2. Analysis, simulation, and identification of bacterial regulatory networks

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.

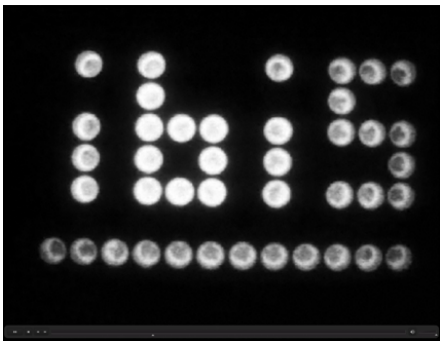


Figure 4. Playful illustration of the principle of reporter genes (see http://ibis.inrialpes.fr for the corresponding movie). A microplate containing a minimal medium (with glucose and acetate) is filmed during 36 hours. Wells contain E. coli bacteria which are transformed with a reporter plasmid containing 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.

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.

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

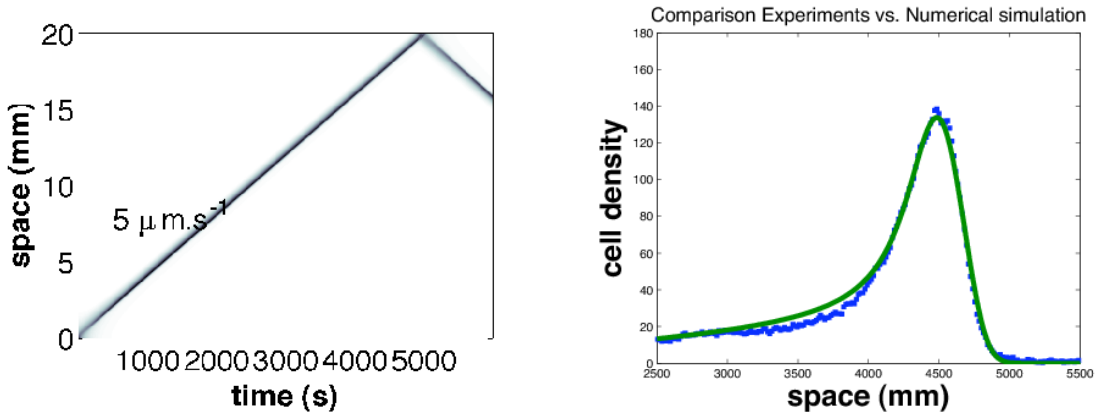


Figure 1. (left) Numerical simulation of a traveling pulse obtained with the kinetic model (right) Comparison between the bacteria density measured experimentally (blue dots) and the density computed from the kinetic model.

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

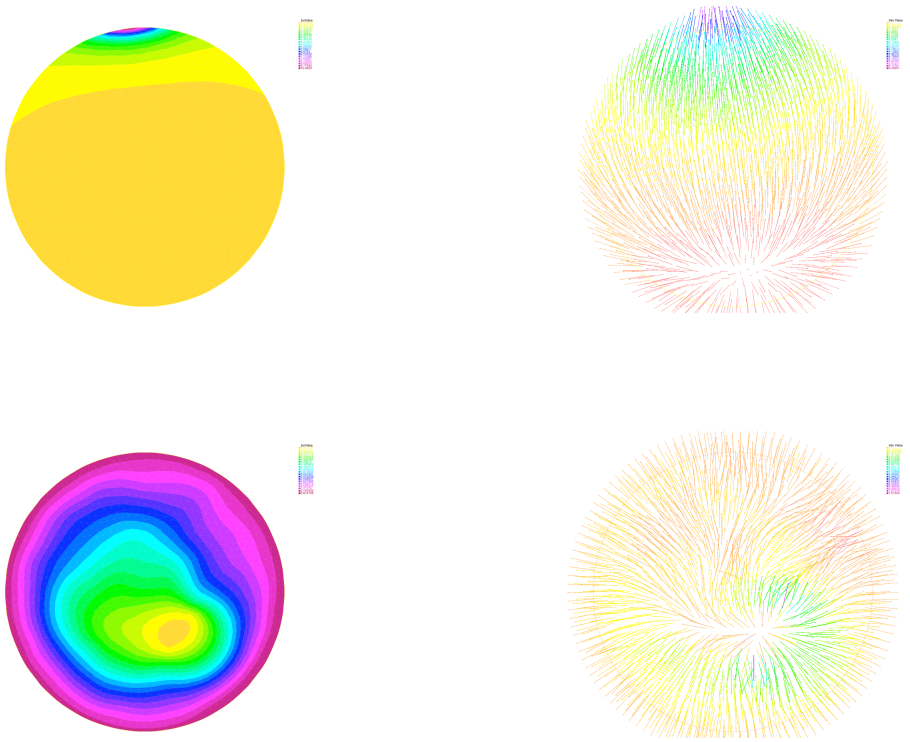


Figure 2. 2D numerical simulations of cell polarization on a round shaped cell. (Top) The actin network carries the attractive field: polarization occurs. (Bottom) The microtubules carry the attractive field: we observe no polarization. (Work in progress; simulations are done with FreeFEM++)

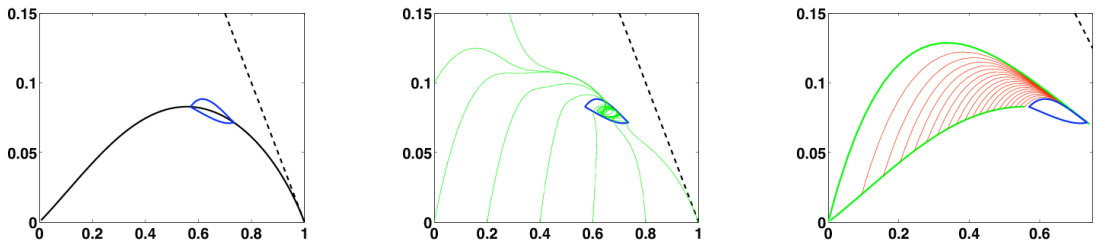


Figure 3. Dynamics of trajectories of the control system projected on the simplex. (left) Remarkable sets in the simplex: the line of eigenvectors parametrized by the control parameter, and the small "ergodic" set. (middle) All trajectories eventually enter the ergodic set. (right) We prove a tunelling 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).

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.

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.

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.

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.

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.

ROMA Team (section vide)

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.

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.

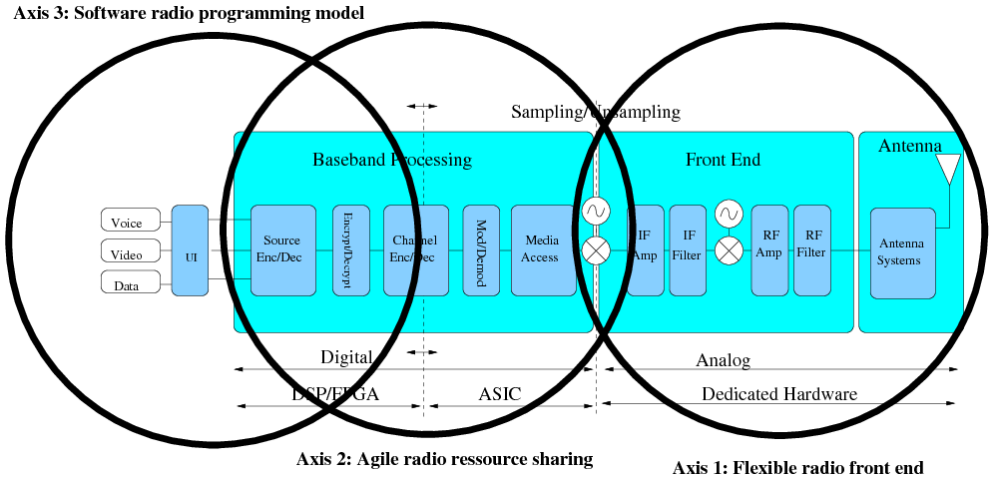


Figure 3. Center of interest for each ot the three Socrate researche axes with respect to a generic software radio terminal.

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

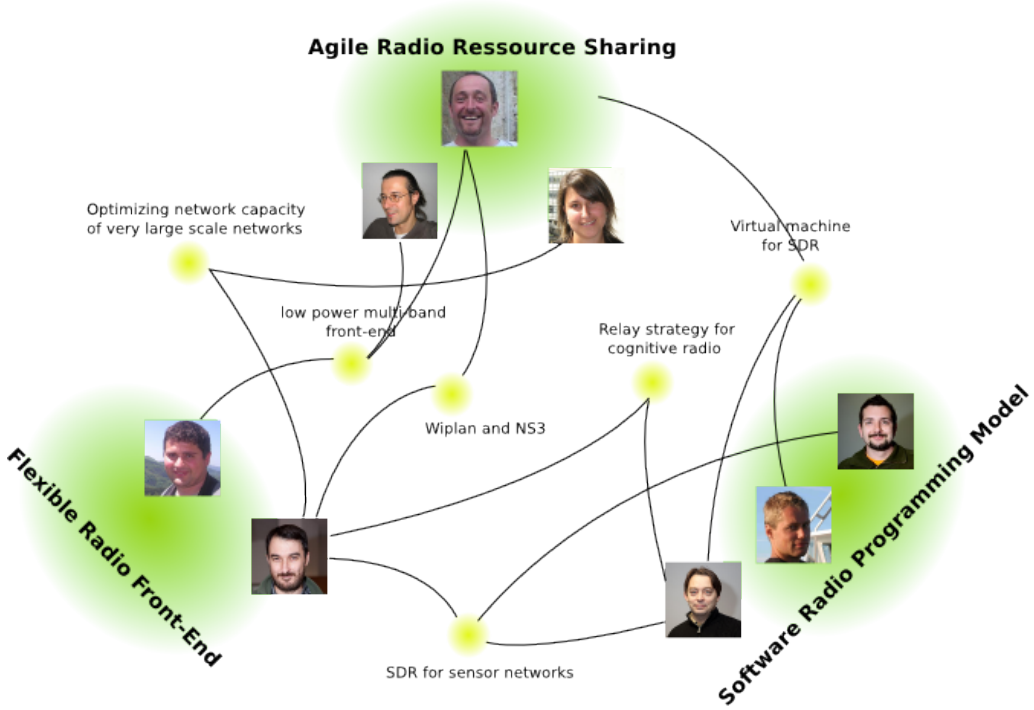


Figure 4. Inter-Axis Collaboration in Socrate: we expect innovative results to come from this pluri-disciplinary research

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.

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.

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.

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.

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.

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.

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.

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.

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 .

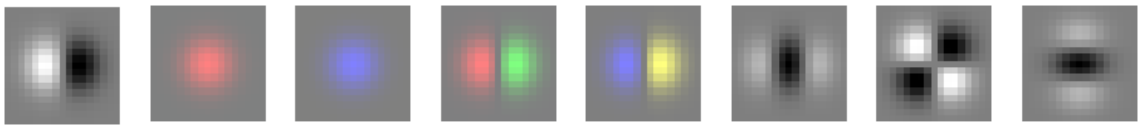


Figure 1. Chromatic Gaussian Receptive Fields ($G_x^L, G^{C_1}, G^{C_2}, G_x^{C_1}, G_x^{C_2}, G_{xx}^L, G_{xy}^L, G_{yy}^L$).

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.

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.