RESEARCH CENTER
Rennes - Bretagne-Atlantique

FIELD

Activity Report 2013

Section New Results

Edition: 2014-03-19

ACES Project-Team

5. New Results

5.1. Self-describing objects

Participants: Michel Banâtre, Nebil Ben Mabrouk, Paul Couderc [contact], Yann Glouche, Arnab Sinha.

Coupled objects enable basic integrity checking for physical objects, and use cases were demonstrated for security and logistics applications. In these applications, high reliability in the RFID reading infrastructure is assumed for the system to work. This suggest another idea for coupled objects: using control data structures distributed over the physical objects in order to improve the reliability of RFID reading protocols. This is the purpose of the Pervasive_RFID project, in collaboration with the IETR which is described in more details below 7.1.2 .

Another development in the line of the coupled objects principles are self-describing objects. While previous works enabled integrity checking over a set of physical objects, these mechanisms were limited in two aspects: expressiveness and autonomy. More precisely, coupled objects support the detection of special conditions (such as a missing element), but not the characterization of these conditions (such as describing the problem, identifying the missing element). Moreover, this compromises the autonomous feature of coupled objects, which would depend on external systems for analyzing these special conditions. Self-describing objects are an attempt to overcome these limitations, and to broaden the application perspectives of autonomous RFID systems.

The principle is to implement distributed data structure over a set of RFID tags, enabling a complex object (made of various parts) or a set of objects belonging to a given logical group to "self-describe" itself and the relation between the various physical elements. Some applications examples includes waste management, assembling and repair assistance, prevention of hazards in situations where various products / materials are combined etc. The key property of self-describing objects is, like for coupled objects, that the vital data are self-"hosted" by the physical element themselves (typically in RFID chips), not an external infrastructure like most RFID systems. This property provides the same advantages as in coupled objects, namely high scalability, easy deployment (no interoperability dependence/interference), and limited risk for privacy.

However, given the extreme storage limitation of RFID chips, designing such systems is difficult:

- data structures must be very frugal in terms of space requirements, both for the structure and for the coding.
- Data structures must be robust and able to survive missing or corrupted elements if we want to ensure the self-describing property for a damaged or incorrect object.

An application of self-describing objects has been proposed in for waste management, in the context of the bin that think project 7.1.1 . A generic graph structure applicable to RFID systems for supporting self-describing objects is proposed in Arnab Sinha's thesis document (to be defended in April 2014).

5.2. Pervasive support for Smart Homes

Participants: Andrey Boytsov, Michele Dominici, Bastien Pietropaoli, Sylvain Roche, Frederic Weis [contact].

A smart home is a residence equipped with information-and-communication-technology (ICT) devices conceived to collaborate in order to anticipate and respond to the needs of the occupants, working to promote their comfort, convenience, security and entertainment while preserving their natural interaction with the environment.

The idea of using the Ubiquitous Computing paradigm in the smart home domain is not new. However, the state-of-the-art solutions only partially adhere to its principles. Often the adopted approach consists in a heavy deployment of sensor nodes, which continuously send a lot of data to a central elaboration unit, in charge of the difficult task of extrapolating meaningful information using complex techniques. This is a *logical approach*. ACES proposed instead the adoption of a *physical approach*, in which the information is spread in the environment, carried by the entities themselves, and the elaboration is directly executed by these entities "inside" the physical space. This allows performing meaningful exchanges of data that will thereafter need a less complicate processing compared to the current solutions. The result is a smart home that can, in an easier and better way, integrate the context in its functioning and thus seamlessly deliver more useful and effective user services. Our contribution aims at implementing the physical approach in a domestic environment, showing a solution for improving both comfort and energy savings.

5.2.1. A multi-level context computing architecture

Computing context is a major subject of interest in smart spaces such as smart homes. Contextual data are necessary for services to adapt themselves to the context and to be as efficient as possible. Contextual data may be obtained via augmented appliances capable of communicating their state and a bunch of sensors. It becomes more and more real with the development of the Internet of Things. Unfortunately, the gathered data are not always directly usable to understand what is going on and to build services on them. In order to address this issue, we studied a multi-level context computing architecture divided in four layers:

- *Exploitation layer*: the highest layer, it exploits con- textual data to provide adapted services
- *Context and situation identification layer*: this is what analyzes ongoing situations and potentially predicts future situations
- *Perception layer*: it offers a first layer of abstraction for small pieces of context independent of deployed sensors
- *Sensing layer*: it mainly consists of the data gathered by sensors

In this architecture, every layer is based on the results of its underlying layers. In 2013, we studied several methods that enable the building of such levels of abstractions (see figure 2). The first level of abstraction coming to mind when describing what people are doing in a Home is high level abstractions such as "cooking". Those activities are then the highest level abstraction we want our system to be able to identify.

We proposed to use plan recognition algorithms to analyze sequences of actions and thus predict future actions of users. It is, in our case, adapted to identify ongoing activities and predict future ones. There exist different plan recognition algorithms. However, one interested us particularly, PHATT introduced by Goldman, Geib and Miller. In order to understand how PHATT is working, it is important to understand the hierarchical task network (HTN) planning problem which is "inverted" by the algorithm to perform plan recognition. It consists in automatically generating a plan starting from a set of tasks to execute and some constraints. In our case, we are able to predict future situations depending of the previously observed situations. To give an example, if we want to predict that the situation dinner will occur soon, it is sufficient to have observed situations such as cooking and/or setting the table. The performances of PHATT have been evaluated by Andrey Andrey Boytstov and Frédéric Weis. These results will be published in 2014.

5.2.2. Propagation of BFT

Context-aware applications have to sense the environment in order to adapt themselves and provide with contextual services. This is the case of Smart Homes equipped with sensors and augmented appliances. However, sensors can be numerous, heterogeneous and unreliable. Thus the data fusion is complex and requires a solid theory to handle those problems. For this purpose, we adopted the belief functions theory (BFT). The aim of the data fusion, in our case, is to compute small pieces of context we call context attributes. Those context attributes are diverse and could be for example the presence in a room, the number of people in a room or even that someone may be sleeping in a room. Since the BFT requires a substantial amount of computations, we proposed to reduce as much as possible the number of evidence required to compute a context attribute. Moreover, the number of possible worlds, *i.e.* the number of possible states for a context attribute, is also an

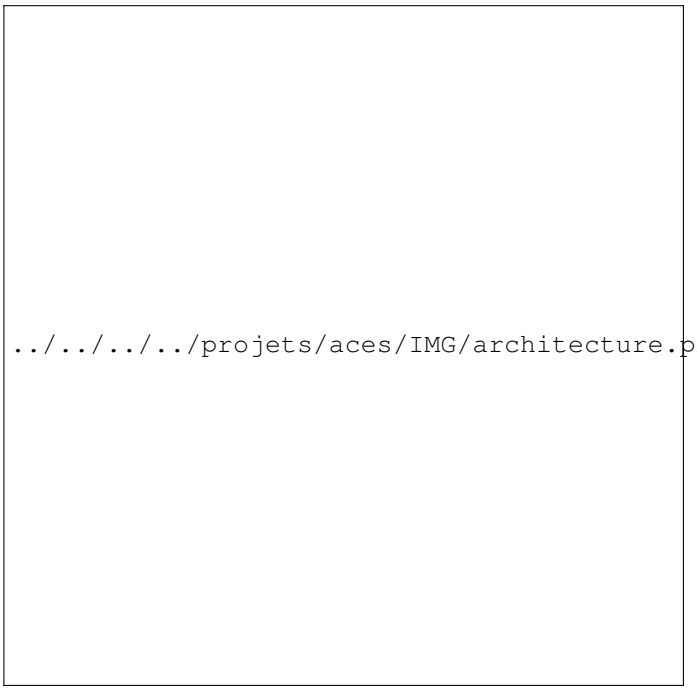


Figure 2. Multi-level context computing architecture

important source of computation. Thus, reducing the number of possible worlds we are working on is also important.

It is especially problematic when working on embedded systems, which may be the case when trying to observe context in smart homes. Thus, with this objective in mind, we observed that some context attributes could be used to compute others. By doing this, the number of gathered and combined evidence for each context attribute could be drastically reduced. This principle is illustrated by Figure 3 : the sets of possible worlds for "Presence" and "Posture" are seen as subsets of "Sleeping". So we proposed and implemented a method to propagate BFT through a set of possible states for a context attribute.

Figure 3. Propagation of Belief Functions Theories

5.2.3. Definition of virtual sensors

In our multi-level architecture, the sensor measures may be imperfect for multiple reasons. The most annoying reasons when deploying a system are biases and noisy measures. It requires fine tuning each type the system is deployed in a new environment. In order to prevent from doing this work again and again at levels where models are hard to build, we proposed to add a new sublayer to the sensing layer (see Figure 2): virtual sensors. Instead of modifying high level models, we created sensor abstractions such as motion sensor, sound sensor, temperature sensor, etc. It is particularly convenient when working with typed data such as temperature or sound level. It is possible to use different brands of sensors for sensors of the same type. Thus, those sensors, even if they are measuring the same physical event, can return very different data due to their range, sensibility, voltage, etc. By creating abstraction of sensors, it is possible to build models directly from typed data simplifying even more the building of models as those data have are understandable by humans. Those virtual sensors are built very simply from common heuristics and can be used for ias and noise compensation, Data aggregation and Meta-data generation.

It is also possible in these virtual sensors to implement fault and failure detection mechanisms using the BFT. It enables the detection of fault in the case of sensors of the same type. At higher level, those mechanisms will detect inconsistency between sensors of different types which is not of the same utility. Thus, those virtual virtual sensors, without disabling any features in our architecture, bring more stability for our models. Moreover, by keeping the virtual sensors very simple, they are easy to adapt and tune in a new environment and the overhead in terms of computation is reduced to the minimum and does not really impact the global system performance. Finally, the fine tuning part is always reduced to this level of our architecture and nothing else has to be changed when we move the system from one environment to another.

ALF Project-Team

6. New Results

6.1. Processor Architecture within the ERC DAL project

Participants: Pierre Michaud, Nathanaël Prémillieu, Luis Germán Garcia Morales, Bharath Narasimha Swamy, Sylvain Collange, André Seznec, Arthur Perais, Surya Natarajan, Sajith Kalathingal, Tao Sun, Andrea Mondelli, Aswinkumar Sridharan, Alain Ketterlin, Kamil Kedzierski.

Processor, cache, locality, memory hierarchy, branch prediction, multicore, power, temperature

Multicore processors have now become mainstream for both general-purpose and embedded computing. Instead of working on improving the architecture of the next generation multicore, with the DAL project, we deliberately anticipate the next few generations of multicores. While multicores featuring 1000s of cores might become feasible around 2020, there are strong indications that sequential programming style will continue to be dominant. Even future mainstream parallel applications will exhibit large sequential sections. Amdahl's law indicates that high performance on these sequential sections is needed to enable overall high performance on the whole application. On many (most) applications, the effective performance of future computer systems using a 1000-core processor chip will significantly depend on their performance on both sequential code sections and single threads.

We envision that, around 2020, the processor chips will feature a few complex cores and many (may be 1000's) simpler, more silicon and power effective cores.

In the DAL research project, http://www.irisa.fr/alf/index.php?option=com_content&view=article&id=55&Itemid=3&lang=en, we explore the microarchitecture techniques that will be needed to enable high performance on such heterogeneous processor chips. Very high performance will be required on both sequential sections, -legacy sequential codes, sequential sections of parallel applications-, and critical threads on parallel applications, -e.g. the main thread controlling the application. Our research focuses essentially on enhancing single process performance.

6.1.1. Microarchitecture exploration of control flow reconvergence

Participants: Nathanaël Prémillieu, André Seznec.

After continuous progress over the past 15 years [8], [10], the accuracy of branch predictors seems to be reaching a plateau. Other techniques to limit control dependency impact are needed. Control flow reconvergence is an interesting property of programs. After a multi-option control-flow instruction (i.e. either a conditional branch or an indirect jump including returns), all the possible paths merge at a given program point: the reconvergence point.

Superscalar processors rely on aggressive branch prediction, out-of-order execution and instruction level parallelism for achieving high performance. Therefore, on a superscalar core , the overall speculative execution after the mispredicted branch is cancelled, leading to a substantial waste of potential performance. However, deep pipelines and out-of-order execution induce that, when a branch misprediction is resolved, instructions following the reconvergence point have already been fetched, decoded and sometimes executed. While some of this executed work has to be cancelled since data dependencies exist, canceling the control independent work is a waste of resources and performance. We have proposed a new hardware mechanism called SYRANT, SYmmetric Resource Allocation on Not-taken and Taken paths, addressing control flow reconvergence at a reasonable cost. Moreover, as a side contribution of this research we have shown that, for a modest hardware cost, the outcomes of the branches executed on the wrong paths can be used to guide branch prediction on the correct path [13].

6.1.2. Efficient Execution on Guarded Instruction Sets

Participants: Nathanaël Prémillieu, André Seznec.

ARM ISA based processors are no longer low complexity processors. Nowadays, ARM ISA based processor manufacturers are struggling to implement medium-end to high-end processor cores which implies implementing a state-of-the-art out-of-order execution engine. Unfortunately providing efficient out-of-order execution on legacy ARM codes may be quite challenging due to guarded instructions.

Predicting the guarded instructions addresses the main serialization impact associated with guarded instructions execution and the multiple definition problem. Moreover, guard prediction allows to use a global branch-and-guard history predictor to predict both branches and guards, often improving branch prediction accuracy. Unfortunately such a global branch-and-guard history predictor requires the systematic use of guard predictions. In that case, poor guard prediction accuracy would lead to poor overall performance on some applications.

Building on top of recent advances in branch prediction and confidence estimation, we propose a hybrid branch and guard predictor, combining a global branch history component and global branch-and-guard history component. The potential gain or loss due to the systematic use of guard prediction is dynamically evaluated at run-time. Two computing modes are enabled: systematic guard prediction use and high confidence only guard prediction use. Our experiments show that on most applications, an overwhelming majority of guarded instructions are predicted. Therefore a relatively inefficient but simple hardware solution can be used to execute the few unpredicted guarded instructions. Significant performance benefits are observed on most applications while applications with poorly predictable guards do not suffer from performance loss [35], [34], [13].

6.1.3. Revisiting Value Prediction

Participants: Arthur Perais, André Seznec.

Value prediction was proposed in the mid 90's to enhance the performance of high-end microprocessors. The research on Value Prediction techniques almost vanished in the early 2000's as it was more effective to increase the number of cores than to dedicate some silicon area to Value Prediction. However high end processor chips currently feature 8-16 high-end cores and the technology will allow to implement 50-100 of such cores on a single die in a foreseeable future. Amdahl's law suggests that the performance of most workloads will not scale to that level. Therefore, dedicating more silicon area to value prediction in high-end cores might be considered as worthwhile for future multicores.

First, we introduce a new value predictor VTAGE harnessing the global branch history [32]. VTAGE directly inherits the structure of the indirect jump predictor ITTAGE [8]. VTAGE is able to predict with a very high accuracy many values that were not correctly predicted by previously proposed predictors, such as the FCM predictor and the stride predictor. Three sources of information can be harnessed by these predictors: the global branch history, the differences of successive values and the local history of values. Moreover, VTAGE does not suffer from short critical prediction loops and can seamlessly handle back-to-back predictions, contrarily to previously proposed, hard to implement FCM predictors.

Second, we show that all predictors are amenable to very high accuracy at the cost of some loss on prediction coverage [32]. This greatly diminishes the number of value mispredictions and allows to delay validation until commit-time. As such, no complexity is added in the out-of-order engine because of VP (save for ports on the register file) and pipeline squashing at commit-time can be used to recover. This is crucial as adding *selective replay* in the OoO core would tremendously increase complexity.

Third, we leverage the possibility of validating predictions at commit to introduce a new microarchitecture, EOLE [31]. EOLE features *Early Execution* to execute simple instructions whose operands are ready in parallel with Rename and *Late Execution* to execute simple predicted instructions and high confidence branches just before Commit. EOLE depends on Value Prediction to provide operands for *Early Execution* and predicted instructions for *Late Execution*. However, Value Prediction requires EOLE to become truly practical. That is, EOLE allows to reduce the out-of-order issue-width by 33% without impeding performance. As such, the number of ports on the register file diminishes. Furthermore, optimizations of the register file such as *banking* further reduce the number of required ports. Overall EOLE possesses a register file whose complexity is on-par with that of a regular wider-issue superscalar while the out-of-order components (scheduler , bypass)

are greatly simplified. Moreover, thanks to Value Prediction, speedup is obtained on many benchmarks of the SPEC'00/'06 suite.

6.1.4. Helper threads

Participants: Bharath Narasimha Swamy, Alain Ketterlin, André Seznec.

As the number of cores on die increases with the improvements in silicon process technology, the strategy of replicating identical cores does not scale to meet the performance needs of mixed workloads. Heterogeneous Many Cores (HMC) that mix many simple cores with a few complex cores are emerging as a design alternative that can provide both high performance and power-efficient execution. The availability of many simple cores in a HMC presents an opportunity to utilize low power cores to accelerate sequential execution on the complex core. For example simple cores can execute pre-computational (or helper) code and generate prefetch requests for the main thread.

We explore the design of a lightweight architectural framework that provides instruction set support and a low-latency interface to simple-cores for efficient helper code execution. We utilize static analyses and profile data to generate helper codelets that target delinquent loads in the main thread. The main thread is instrumented to initiate helper execution ahead of time, and utilizes instruction set support to signal helper execution on the simple core, and to pass live-in values for the helper codelet. Pre-computational code executes on the simple core and generates prefetch requests that install data into a shared last-level cache. Initial experiments with a trace based simulation framework show that helper execution has the potential to cover cache-missing loads on the main thread.

The restriction of prefetching to a lower level shared cache in a loosely coupled system limits the benefits of helper execution. The main thread should have a low latency access mechanism to data prefetched by helper execution. We plan to explore direct, yet light weight, mechanisms for data communication between the helper core and the main core.

6.1.5. Adaptive Intelligent Memory Systems

Participants: André Seznec, Aswinkumar Sridharan.

On multicores, the processors are sharing the memory hierarchy, buses, caches, and memory. The performance of any single application is impacted by its environment and the behavior of the other applications co-running on the multicore. Different strategies have been proposed to isolate the behavior of the different co-running applications, for example performance isolation cache partitioning, while several studies have addressed the global issue of optimizing throughput through the cache management.

However these studies are limited to a few cores (2-4-8) and generally features mechanisms that cannot scale to 50-100 cores. Moreover so far the academic propositions have generally taken into account a single parameter, the cache replacement policy or the cache partitioning. Other parameters such as cache prefetching and its aggressiveness already impact the behavior of a single thread application on a uniprocessor. Cache prefetching policy of each thread will also impact the behavior of all the co-running threads.

Our objective is to define an Adaptive and Intelligent Memory System management hardware, AIMS. The goal of AIMS will be to dynamically adapt the different parameters of the memory hierarchy access for each individual co-running process in order to achieve a global objective such as optimized throughput, thread fairness or respecting quality of services for some privileged threads.

6.1.6. Modeling multi-threaded programs execution time in the many-core era

Participants: Surya Natarajan, Bharath Narasimha Swamy, André Seznec.

Multi-core have become ubiquitous and industry is already moving towards the many-core era. Many open-ended questions remain unanswered for the upcoming many-core era. From the software perspective, it is unclear which applications will be able to benefit from many cores. From the hardware perspective, the tradeoff between implementing many simple cores, fewer medium aggressive cores or even only a moderate number of aggressive cores is still in debate. Estimating the potential performance of future parallel applications on the yet-to-be-designed future many cores is very speculative. The simple models proposed by Amdahl's law or Gustafson's law are not sufficient and may lead to erroneous conclusions. In this paper, we propose a still simple execution time model for parallel applications, the SNAS model. As previous models, the SNAS model evaluates the execution time of both the serial part and the parallel part of the application, but takes into account the scaling of both these execution times with the input problem size and the number of processors. For a given application, a few parameters are collected on the effective execution of the application with a few threads and small input sets. The SNAS model allows to extrapolate the behavior of a future application exhibiting similar scaling characteristics on a many core and/or a large input set. Our study shows that the execution time of the serial part of many parallel applications tends to increase along with the problem size, and in some cases with the number of processors. It also shows that the efficiency of the execution of the parallel part decreases dramatically with the number of processors for some applications. Our model also indicates that since several different applications scaling will be encountered, hybrid architectures featuring a few aggressive cores and many simple cores should be privileged.

6.1.6.1. Augmenting superscalar architecture for efficient many-thread parallel execution

Participants: Sylvain Collange, André Seznec, Sajith Kalathingal.

We aim at exploring the design of a unique core that efficiently run both sequential and massively parallel sections. We explore how the architecture of a complex superscalar core has to be modified or enhanced to be able to support the parallel execution of many threads from the same application (10's or even 100's a la GPGPU on a single core).

SIMD execution is the preferred way to increase energy efficiency on data-parallel workloads. However, explicit SIMD instructions demand challenging auto-vectorization or manual coding, and any change in SIMD width requires at least a recompile, and typically manual code changes. Rather than vectorize at compile-time, our approach is to dynamically vectorize SPMD programs at the micro-architectural level. The SMT-SIMD hybrid core we propose extracts data parallelism from thread parallelism by scheduling groups of threads in lockstep, in a way inspired by the execution model of GPUs. As in GPUs, conditional branches whose outcome differ between threads are handled with conditionally masked execution. However, while GPUs rely on explicit re-convergence instructions to restore lockstep execution, we target existing general-purpose instruction sets, in order to run legacy binary programs. Thus, the main challenge consists in detecting re-convergence points dynamically.

We proposed instruction fetch policies that apply heuristics to maximize the cycles spent in lockstep execution. We evaluated their efficiency and performance impact on an out-of-order superscalar core simulator. Results validate the viability of our approach, by showing that existing compiled SPMD programs are amenable to lockstep execution without modification nor recompilation.

6.2. Other Architecture Studies

Participants: Damien Hardy, Pierre Michaud, Ricardo Andrés Velásquez, Sylvain Collange, André Seznec, Sajith Kalathingal, Junjie Lai.

GPU, performance, simulation, vulnerability

6.2.1. Performance Upperbound Analysis of GPU applications

Participants: Junjie Lai, André Seznec.

In the framework of the ANR Cosinus PetaQCD project (ended Oct 2012), we have been modeling the demands of high performance scientific applications on hardware. GPUs have become popular and cost-effective hardware platforms. In this context, we have been addressing the gap between theoretical peak performance on GPU and the effective performance. There have been many studies on optimizing specific applications on GPU and also a lot of studies on automatic tuning tools. However, the gap between the effective performance and the maximum theoretical performance is often huge. A tighter performance upperbound of an application is needed in order to evaluate whether further optimization is worth the effort. We designed a new approach to compute the CUDA application's performance upperbound through intrinsic algorithm information coupled with low-level hardware benchmarking. Our analysis [11], [22] allows us to understand which parameters are critical to the performance and have more insights of the performance result. As an example, we analyzed the performance upperbound of SGEMM (Single-precision General Matrix Multiply) on Fermi and Kepler GPUs. Through this study, we uncover some undocumented features on Kepler GPU architecture. Based on our analysis, our implementations of SGEMM achieve the best performance on Fermi and Kepler GPUs so far (5 % improvement on average).

6.3. Microarchitecture Performance Analysis

Participants: Ricardo Andrés Velásquez, Pierre Michaud, André Seznec.

6.3.1. Selecting benchmark combinations for the evaluation of multicore throughput

Participants: Ricardo Andrés Velásquez, Pierre Michaud, André Seznec.

In [26], we have shown that fast approximate microarchitecture models such as BADCO [16] can be useful for selecting multiprogrammed workloads for evaluating the throughput of multicore processors. Computer architects usually study multiprogrammed workloads by considering a set of benchmarks and some combinations of these benchmarks. However, there is no standard method for selecting such sample, and different authors have used different methods. The choice of a particular sample impacts the conclusions of a study. Using BADCO, we propose and compare different sampling methods for defining multiprogrammed workloads for computer architecture. We evaluate their effectiveness on a case study, the comparison of several multicore last-level cache replacement policies. We show that random sampling, the simplest method, is robust to define a representative sample of workloads, provided the sample is big enough. We propose a method for estimating the required sample size based on fast approximate simulation. We propose a new method, workload stratification, which is very effective at reducing the sample size in situations where random sampling would require large samples.

6.3.2. A systematic approach for defining multicore throughput metrics

Participant: Pierre Michaud.

This research was done in collaboration with Stijn Eyerman from Ghent University.

Measuring throughput is not as straightforward as measuring execution time. This has led to an ongoing debate on what forms a meaningful throughput metric for multi-program workloads. In [29], we present a method to construct throughput metrics in a systematic way: we start by expressing assumptions on job size, job distribution, scheduling, etc., that together define a theoretical throughput experiment. The throughput metric is then the average throughput of this experiment. Different assumptions lead to different metrics, so one should select the metric whose assumptions are close to the real usage he/she has in mind. We elaborate multiple metrics based on different assumptions. In particular, we identify the assumptions that lead to the commonly used weighted speedup and harmonic mean of speedups. Our study clarifies that they are actual throughput metrics, which was recently questioned. We also propose some new throughput metrics, whose calculation sometimes requires approximation. We use synthetic and real experimental data to characterize metrics and show how they relate to each other. Our study can also serve as a starting point if one needs to define a new metric based on specific assumptions, other than the ones we consider in this study. Throughput metrics should always be defined from explicit assumptions, because this leads to a better understanding of the implications and limits of the results obtained with that metric.

6.4. Compiler, vectorization, interpretation

Participants: Erven Rohou, Emmanuel Riou, Arjun Suresh, André Seznec, Nabil Hallou, Alain Ketterlin, Sylvain Collange.

6.4.1. Vectorization Technology To Improve Interpreter Performance

Participant: Erven Rohou.

Recent trends in consumer electronics have created a new category of portable, lightweight software applications. Typically, these applications have fast development cycles and short life spans. They run on a wide range of systems and are deployed in a target independent bytecode format over Internet and cellular networks. Their authors are untrusted third-party vendors, and they are executed in secure managed runtimes or virtual machines. Furthermore, due to security policies, these virtual machines are often lacking just-in-time compilers and are reliant on interpreter execution.

The main performance penalty in interpreters arises from instruction dispatch. Each bytecode requires a minimum number of machine instructions to be executed. In this work we introduce a powerful and portable representation that reduces instruction dispatch thanks to vectorization technology. It takes advantage of the vast research in vectorization and its presence in modern compilers. Thanks to a split compilation strategy, our approach exhibits almost no overhead. Complex compiler analyses are performed ahead of time. Their results are encoded on top of the bytecode language, becoming new SIMD IR (i.e., intermediate representation) instructions. The bytecode language remains unmodified, thus this representation is compatible with legacy interpreters.

This approach drastically reduces the number of instructions to interpret and improves execution time. [15]. SIMD IR instructions are mapped to hardware SIMD instructions when available, with a substantial improvement.

6.4.2. Improving sequential performance: the case of floating point computations

Participants: Erven Rohou, André Seznec, Arjun Suresh.

One way to enhance sequential performance is to consider floating point computations. Languages and instruction sets provide support for only a few representations, namely float and double, and programmers are likely to use the most accurate (unless they handle large data structures). Still, in most cases, programmers do not formally specify the precision they require from their applications, and have no guarantee on the precision they actually get. This is an opportunity for a tradeoff between performance and precision: programs could run faster at the expense of a less accurate result (note that existing compilers already embed some unsafe transformations, for example when flags such as -fast or -ffastmath are used).

The first step consisted in applying memoization to the math library libm. In this case, results are still correct. The performance improvement comes from caching results of pure functions, and retrieving them instead of recomputing a result. This shows good results on floating point intensive benchmarks. In a next step, a helper thread will monitor the patterns of parameters and precompute likely values to "prefetch" results ahead of time.

Reduced precision comes into play when no pattern can be identified, but the new value is close enough to already computed values. We plan to apply interpolation to compute the result faster than the standard code. We will also investigate how we can leverage known properties of mathematical functions, as well as programmer hints about useful properties of user-defined functions, and where reduced precision is acceptable.

6.4.3. Identifying divergence in GPU architectures

Participant: Sylvain Collange.

This research is done in collaboration with Fernando M. Q. Pereira, Diogo Sampaio and Rafael Martins de Souza, UFMG, Brazil.

GPU architectures rely on SIMD execution by vectorizing across SPMD threads. They achieve the best performance when consecutive threads take the same paths through conditional branches and access contiguous memory locations. Thus, many GPU code optimizations that target the control flow or memory access patterns necessitate accurate information about which branches and memory accesses are divergent across threads.

To enable such optimizations, we proposed divergence analysis, a compiler pass that identifies similarities in the control flow and data flow of concurrent threads [37]. This static analysis identifies program variables that are affine functions of the thread identifier and propagate this knowledge to conditional branches and memory accesses. Our analysis consistently outperforms other comparable analyses, thanks to the combination of taking into account affine relations between variables and accurately modeling control dependencies.

6.4.4. Code Obfuscation

Participant: Erven Rohou.

This research is done in collaboration with the group of Prof. Ahmed El-Mahdy at E-JUST, Alexandria, Egypt.

We proposed to leverage JIT compilation to make software tamper-proof. The idea is to constantly generate different versions of an application, even while it runs, to make reverse engineering hopeless. More precisely a JIT engine is used to generate new versions of a function each time it is invoked, applying different optimizations, heuristics and parameters to generate diverse binary code. A strong random number generator will guarantee that generated code is not reproducible, though the functionality is the same [38].

On-Stack-Replacement has been previously proposed to recompile functions while they run. However, it relies on compiler-generated switch points. We proposed a new technique to recompile functions at arbitraty points, thus reinforcing the Obfuscating JIT approach. A prototype is being developed [27].

A new obfuscation technique based of decomposition of CFGs into threads has been proposed. We exploit the mainstream multi-core processing in these systems to substantially increase the complexity of programs, making reverse engineering more complicated. The novel method automatically partitions any serial thread into an arbitrary number of parallel threads, at the basic-block level. The method generates new control-flow graphs, preserving the blocks' serial successor relations and guaranteeing that one basic-block is active at a time through using guards. The method generates m^n different combinations for m threads and n basic-blocks, significantly complicating the execution state. We also provide proof of correctness for the method.

6.4.5. Padrone

Participants: Erven Rohou, Alain Ketterlin, Emmanuel Riou.

The objective of the ADT PADRONE is to design and develop a platform for re-optimization of binary executables at run-time. Development is ongoing, and an early prototype is functional. In [24], we described the infrastructure of Padrone, and showed that its profiling overhead is minimum. We illustrated its use through two examples. The first example shows how a user can easily write a tool to identify hotspots in their application, and how well they perform (for example, by computing the number of executed instructions per cycle). In the second example, we illustrate the replacement of a given function (typically a hotspot) by an optimized version, while the program runs.

We believe PADRONE fills an empty design point in the ecosystem of dynamic binary tools.

6.4.6. Dynamic Analysis and Re-Optimization

Participants: Erven Rohou, Emmanuel Riou, Nabil Hallou, Alain Ketterlin.

This work is done in collaboration with Philippe Clauss (Inria CAMUS).

Dynamic binary analysis and re-optimization is specially interesting for legacy or commercial applications, but also in the context of cloud deployment, where actual hardware is unknown, and other applications competing for hardware resources can vary.

Initial results show that we are able to identify function hotspots that contain vectorized code for the Intel SSE extension, analyze them, and reoptimize the loops to target the latest and more powerful AVX ISA extension.

6.4.7. Branch Prediction and Performance of Interpreter

Participants: Erven Rohou, André Seznec, Bharath Narasimha Swamy.

Interpreters have been used in many contexts. They provide portability and ease of development at the expense of performance. The literature of the past decade covers analysis of why interpreters are slow, and many software techniques to improve them. A large proportion of these works focuses on the dispatch loop, and in particular on the implementation of the switch statement: typically an indirect branch instruction. Conventional wisdom attributes a significant penalty to this branch, due to its high misprediction rate. We revisit this assumption [36], considering current interpreters, and modern predictors. Using both hardware counters and simulation, we show that the accuracy of indirect branch prediction is no longer critical for interpreters. We also compare the characteristics of these interpreters and analyze why the indirect branch is less important than before.

6.5. WCET estimation

Participants: Damien Hardy, Benjamin Lesage, Hanbing Li, Isabelle Puaut, Erven Rohou, André Seznec.

Predicting the amount of resources required by embedded software is of prime importance for verifying that the system will fulfill its real-time and resource constraints. A particularly important point in hard real-time embedded systems is to predict the Worst-Case Execution Times (WCETs) of tasks, so that it can be proven that tasks temporal constraints (typically, deadlines) will be met. Our research concerns methods for obtaining automatically upper bounds of the execution times of applications on a given hardware. Our new results this year are on (i) multi-core architectures (ii) WCET estimation for faulty architectures (iii) traceability of flow information in compilers for WCET estimation.

6.5.1. WCET estimation and multi-core systems

6.5.1.1. Predictable shared caches for mixed-criticality real-time systems

Participants: Benjamin Lesage, Isabelle Puaut, André Seznec.

The general adoption of multi-core architectures has raised new opportunities as well as new issues in all application domains. In the context of real-time applications, it has created one major opportunity and one major difficulty. On the one hand, the availability of multiple high performance cores has created the opportunity to mix on the same hardware platform the execution of a complex critical real-time workload and the execution of non-critical applications. On the other hand, for real-time tasks timing deadlines must be met and enforced. Hardware resource sharing inherent to multicores hinders the timing analysis of concurrent tasks. Two different objectives are then pursued: enforcing timing deadlines for real-time tasks and achieving highest possible performance for the non-critical workload.

In this work, we suggest a hybrid hardware-based cache partitioning scheme that aims at achieving these two objectives at the same time. Plainly considering inter-task conflicts on shared cache for real-time tasks yields very pessimistic timing estimates. We remove this pessimism by reserving private cache space for real-time tasks. Upon the creation of a real-time task, our scheme reserves a fixed number of cache lines per set for the task. Therefore uniprocessor worst case execution time (WCET) estimation techniques can be used, resulting in tight WCET estimates. Upon the termination of the real-time task, this private cache space is released and made available for all the executed threads including non-critical ones. That is, apart the private spaces reserved for the real-time tasks currently running, the cache space is shared by all tasks running on the processor, i.e. non-critical tasks but also the real-time tasks for their least recently used blocks. Experiments show that the proposed cache scheme allows to both guarantee the schedulability of a set of real-time tasks with tight timing constraints and enable high performance on the non-critical tasks.

This work is the main contribution of the PhD thesis of Benjamin Lesage [12].

6.5.1.2. WCET estimation for massively parallel processor arrays

Participant: Isabelle Puaut.

This is joint work with Dumitru Potop-Butucaru, Inria, EPI AOSTE.

Classical timing analysis techniques for parallel code isolates micro-architecture analysis from the analysis of synchronizations between cores by performing them in two separate analysis phases (WCET – worst-case execution time – and WCRT – worst-case response time analyses). This isolation has its advantages, such as a reduction of the complexity of each analysis phase, and a separation of concerns that facilitates the development of analysis tools. But isolation also has a major drawback: a loss in precision which can be significant. To consider only one aspect, to be safe the WCET analysis of each synchronization-free sequential code region has to consider an undetermined micro-architecture state. This may result in overestimated WCETs, and consequently on pessimistic execution time bounds for the whole parallel application. The contribution of this work [33], [23] is an *integrated* WCET analysis approach that considers at the same time micro-architectural information and the synchronizations between cores. This is achieved by extending a state-of-the-art WCET estimation technique and tool to manage synchronizations and communications between the sequential threads running on the different cores. The benefits of the proposed method are twofold. On the one hand, the micro-architectural state is not lost between synchronization-free code regions running on the same core, which results in tighter execution time estimates. On the other hand, only one tool is required for the temporal validation of the parallel application, which reduces the complexity of the timing validation toolchain.

Such a holistic approach is made possible by the use of deterministic and composable software and hardware architectures (homogeneous multi-cores without cache sharing, static assignment of the code regions on the cores). We demonstrate the interest of the approach using an adaptive differential pulse-code modulation (*adpcm*) encoder where the integrated WCET approach provides significantly tighter response time estimations than the more classical WCRT approaches, with a gain of 21% on average.

6.5.2. WCET estimation for architectures with faulty caches

Participants: Damien Hardy, Isabelle Puaut.

Semiconductor technology evolution suggests that permanent failure rates will increase dramatically with scaling, in particular for SRAM cells. While well known approaches such as error correcting codes exist to recover from failures and provide fault-free chips, they will not be affordable anymore in the future due to their non-scalable cost. Consequently, other approaches like fine grain disabling and reconfiguration of hardware elements (e.g. individual functional units or cache blocks) will become economically necessary. This fine-grain disabling will lead to degraded performance compared to a fault-free execution.

A common implicit assumption in all static worst-case execution time (WCET) estimation methods is that the hardware is not subject to faults. Their result is not safe anymore when using fine grain disabling of hardware components, which degrades performance.

In [21] a method that statically calculates a probabilistic WCET bound in the presence of permanent faults in instruction caches is provided. The method, from a given program, cache configuration and probability of cell failure, derives a probabilistic WCET bound. The proposed method, because it relies on static analysis, is guaranteed to identify the longest program path, its probabilistic nature only stemming from the presence of faults. The method is computationally tractable because it does not require an exhaustive enumeration of the possible locations of faulty cache blocks. Experimental results show that it provides WCET estimates very close to, but never below, the method that derives probabilistic WCETs by enumerating all possible locations of faulty cache blocks. The proposed method not only allows to quantify the impact of permanent faults on WCET estimates, but also can be used in architectural exploration frameworks to select the most appropriate fault management mechanisms.

6.5.3. Traceability of flow information for WCET estimation

Participants: Hanbing Li, Isabelle Puaut, Erven Rohou.

This research is part of the ANR W-SEPT project.

Control-flow information is mandatory for WCET estimation, to guarantee that programs terminate (e.g. provision of bounds for the number of loop iterations) but also to obtain tight estimates (e.g. identification of infeasible or mutually exclusive paths). Such flow information is expressed though annotations, that may be calculated automatically by program/model analysis, or provided manually.

The objective of this work is to address the challenging issue of the mapping and transformation of the flow information from high level down to machine code. In a first step, we have considered the issue of conveying information through the compilation flow, without any optimization. We have created our own WCET information type and used the annotation files FFX (Flow Fact in XML, provided by IRIT, partner of the W-SEPT project), and applied them to the LLVM compiler framework. We are currently studying the impact of optimizations on the traceability of annotations. We are currently designing a framework for flow fact transformation for a large panel of compiler optimizations.

6.6. HPC and mobile computing

Participant: François Bodin.

We have initiated a research action on the interaction between mobile computing and HPC. We aim at studying data representation linked to parallel programming in heterogeneous systems. In particular, we want to explore energy tradeoffs when changing hardware resources from a light mobile platform to remote execution in a datacenter.

As a test case, we are developing an application for inventorying art pieces in the public domain. This is done in collaboration with University of Rennes 2. This test case is a pluridisplinary collaboration whose goal for University of Rennes 2 is to study how mobile computing can contribute to art studies and dissemination.

6.7. Application-specific number systems

Participant: Sylvain Collange.

This research is done in collaboration with Mark G. Arnold, XLNS Research, USA.

Reconfigurable FPGA platforms let designers build efficient application-specific circuits, when the performance or energy efficiency of general-purpose CPUs is insufficient, and the production volume is not enough to offset the very high cost of building a dedicated integrated circuit (ASIC). One way to take advantage of the flexibility offered by FPGAs is to tailor arithmetic operators for the application. In particular, the Logarithmic Number System (LNS) is suitable for embedded applications dealing with low-precision, high-dynamic range numbers.

Like floating-point, LNS can represent numbers from a wide dynamic range with constant relative accuracy. However, while standard floating-point offer so-called subnormal numbers to represent numbers close to zero with constant absolute accuracy, LNS numbers abruptly overflow to zero, resulting in a gap in representable numbers close to zero that can impact the accuracy of numerical algorithms.

We proposed a generalization of LNS that incorporate features analogous to subnormal floating-point [18], [28]. The Denormal LNS (DLNS) system we introduce defines a class of hybrid number systems that offer quasi-constant absolute accuracy close to zero and quasi-constant relative accuracy on larger numbers. These systems can be configured to range from pure LNS (constant relative accuracy) to fixed-point (constant absolute accuracy across the whole range).

ASAP Project-Team

6. New Results

6.1. Models and abstractions for distributed systems

6.1.1. Randomized loose renaming in O(loglog n) time

Participant: George Giakkoupis.

Renaming is a classic distributed coordination task in which a set of processes must pick distinct identifiers from a small namespace. In [24], we consider the time complexity of this problem when the namespace is linear in the number of participants, a variant known as loose renaming. We give a non-adaptive algorithm with $O(\log\log n)$ (individual) step complexity, where n is a known upper bound on contention, and an adaptive algorithm with step complexity $O((\log\log k)^2)$, where k is the actual contention in the execution. We also present a variant of the adaptive algorithm which requires $O(k\log\log k)$*total* process steps. All upper bounds hold with high probability against a strong adaptive adversary. We complement the algorithms with an $\Omega(\log\log n)$ expected time lower bound on the complexity of randomized renaming using test-and-set operations and linear space. The result is based on a new coupling technique, and is the first to apply to non-adaptive randomized renaming. Since our algorithms use $O(n)$ test-and-set objects, our results provide matching bounds on the cost of loose renaming in this setting.

This work was done in collaboration with Dan Alistarh, James Aspnes, and Philipp Woelfel.

6.1.2. An O(sqrt n) space bound for obstruction-free leader election

Participant: George Giakkoupis.

In [32] we present a deterministic obstruction-free implementation of leader election from $O(\sqrt{n})$ atomic $O(\log n)$-bit registers in the standard asynchronous shared memory system with n processes. We provide also a technique to transform any deterministic obstruction-free algorithm, in which any process can finish if it runs for b steps without interference, into a randomized wait-free algorithm for the oblivious adversary, in which the expected step complexity is polynomial in n and b. This transformation allows us to combine our obstruction-free algorithm with the leader election algorithm by Giakkoupis and Woelfel (2012), to obtain a fast randomized leader election (and thus test-and-set) implementation from $O(\sqrt{n})O(\log n)$-bit registers, that has expected step complexity $O(\log^* n)$ against the oblivious adversary. Our algorithm provides the first sub-linear space upper bound for obstruction-free leader election. A lower bound of $\Omega(\log n)$ has been known since 1989 (Styer and Peterson, 1989). Our research is also motivated by the long-standing open problem whether there is an obstruction-free consensus algorithm which uses fewer than n registers.

This work was done in collaboration with Maryam Helmi, Lisa Higham, and Philipp Woelfel.

6.1.3. Broadcast in recurrent dynamic systems

Participants: Michel Raynal, Julien Stainer.

This work [50] proposes a simple broadcast algorithm suited to dynamic systems where links can repeatedly appear and disappear. The algorithm is proved correct and a simple improvement is introduced, that reduces the number and the size of control messages. As it extends in a simple way a classical network traversal algorithm (due to A. Segall, 1983) to the dynamic context, the proposed algorithm has also pedagogical flavor.

This work has been done in collaboration with Jiannong Cao and Weigang Wu.

6.1.4. Computing in the presence of concurrent solo executions

Participants: Michel Raynal, Julien Stainer.

In a wait-free model any number of processes may crash. A process runs solo when it computes its local output without receiving any information from other processes, either because they crashed or they are too slow. While in wait-free shared-memory models at most one process may run solo in an execution, any number of processes may have to run solo in an asynchronous wait-free message-passing model. This work [47] is on the computability power of models in which several processes may concurrently run solo. We introduced a family of round-based wait-free models, called the d-solo models, $1 \leq d \leq n$, where up to d processes may run solo. Then we gave a characterization of the colorless tasks that can be solved in each d-solo model. We also introduced the (d, ϵ)-solo approximate agreement task, which generalizes ϵ-approximate agreement, and proves that (d, ϵ)-solo approximate agreement can be solved in the d-solo model, but cannot be solved in the $(d + 1)$-solo model. We also studied the relation linking d-set agreement and (d, ϵ)-solo approximate agreement in asynchronous wait-free message-passing systems. These results establish for the first time a hierarchy of wait-free models that, while weaker than the basic read/write model, are nevertheless strong enough to solve non-trivial tasks.

This work was done in collaboration with Maurice Herlihy and Sergio Rajsbaum.

6.1.5. Relating message-adversaries and failure detectors

Participants: Michel Raynal, Julien Stainer.

A message adversary is a daemon that suppresses messages in round-based message-passing synchronous systems in which no process crashes. A property imposed on a message adversary defines a subset of messages that cannot be eliminated by the adversary. It has recently been shown that when a message adversary is constrained by a property denoted TOUR (for tournament), the corresponding synchronous system and the asynchronous crash-prone read/write system have the same computability power for task solvability. In this work [39] we introduced new message adversary properties (denoted SOURCE and QUORUM), and shown that the synchronous round-based systems whose adversaries are constrained by these properties are characterizations of classical asynchronous crash-prone systems (1) in which processes communicate through atomic read/write registers or point-to-point message-passing, and (2) enriched with failure detectors such as Ω and Σ. Hence these properties characterize maximal adversaries, in the sense that they define strongest message adversaries equating classical asynchronous crash-prone systems. They consequently provide strong relations linking round-based synchrony weakened by message adversaries with asynchrony restricted with failure detectors. This not only enriches our understanding of the synchrony/asynchrony duality, but also allows for the establishment of a meaningful hierarchy of property-constrained message adversaries.

6.1.6. A hierarchy of agreement problems from simultaneous consensus to set agreement

Participants: Michel Raynal, Julien Stainer.

In this work [38] we investigated the relation linking the s-simultaneous consensus problem and the k-set agreement problem in wait-free message-passing systems. To this end, we defined the (s, k)-SSA problem which captures jointly both problems: each process proposes a value, executes s simultaneous instances of a k-set agreement algorithm, and has to decide a value so that no more than sk different values are decided. We also introduced a new failure detector class denoted $Z_{s,k}$, which is made up of two components, one focused on the "shared memory object" that allows the processes to cooperate, and the other focused on the liveness of (s, k)-SSA algorithms. A novelty of this failure detector lies in the fact that the definition of its two components are intimately related. We designed a $Z_{s,k}$-based algorithm that solves the (s, k)-SSA problem, and shown that the "shared memory"-oriented part of $Z_{s,k}$ is necessary to solve the (s, k)-SSA problem (this generalizes and refines a previous result that showed that the generalized quorum failure detector Σ_k is necessary to solve k-set agreement). We finally, investigated the structure of the family of (s, k)-SSA problems and introduced generalized (asymmetric) simultaneous set agreement problems in which the parameter k can differ in each underlying k-set agreement instance. Among other points, it shows that, for $s, k > 1$, (a) the $(sk, 1)$-SSA problem is strictly stronger that the (s, k)-SSA problem which is itself strictly stronger than the $(1, ks)$-SSA problem, and (b) there are pairs (s_1, k_1) and (s_2, k_2) such that $s_1k_1 = s_2k_2$ and (s_1, k_1)-SSA and (s_2, k_2)-SSA are incomparable.

6.2. Large-scale and user-centric distributed systems

6.2.1. FreeRec: An anonymous and distributed personalization architecture

Participants: Antoine Boutet, Davide Frey, Arnaud Jégou, Anne-Marie Kermarrec, Heverson Borba Ribeiro.

FreeRec is an anonymous decentralized peer-to-peer architecture designed to bring personalization while protecting the privacy of its users [17], [30], [44]. FreeRec's decentralized approach makes it independent of any entity wishing to collect personal data about users. At the same time, its onion-routing-like gossip-based overlay protocols effectively hide the association between users and their interest profiles without affecting the quality of personalization. The core of FreeRec consists of three layers of overlay protocols: the bottom layer, rps, consists of a standard random peer sampling protocol ensuring connectivity; the middle layer, PRPS, introduces anonymity by hiding users behind anonymous proxy chains, providing mutual anonymity; finally, the top clustering layer identifies for each anonymous user, a set of anonymous nearest neighbors. We demonstrate the effectiveness of FreeRec by building a decentralized and anonymous content dissemination system. Our evaluation by simulation, our PlanetLab experiments, and our probabilistic analysis show that FreeRec effectively decouples users from their profiles without hampering the quality of personalized content delivery.

6.2.2. HyRec: A hybrid recommender system

Participants: Antoine Boutet, Davide Frey, Anne-Marie Kermarrec.

The ever-growing amount of data available on the Internet calls for personalization. Yet, the most effective personalization schemes, such as those based on collaborative filtering (CF), are notoriously resource greedy. HyRec is an online cost-effective scalable system for CF personalization. HyRec relies on a hybrid architecture, offloading CPU-intensive recommendation tasks to front-end client browsers, while retaining storage and orchestration tasks within back-end servers. HyRec has been fully implemented and extensively evaluated on several workloads from MovieLens and Digg. We convey the ability of HyRec to significantly reduce the operation costs of the content provider by up to 70% and drastically improve the scalability by up to 500%, with respect to a centralized (or cloud-based recommender approach), while preserving the quality of the personalization. We also show that HyRec is virtually transparent to the users and induces only 3% of the bandwidth consumption of a P2P solution.

6.2.3. Social market

Participants: Davide Frey, Arnaud Jégou, Anne-Marie Kermarrec, Michel Raynal, Julien Stainer.

The ability to identify people that share one's own interests is one of the most interesting promises of the Web 2.0 driving user-centric applications such as recommendation systems or collaborative marketplaces. To be truly useful, however, information about other users also needs to be associated with some notion of trust. Consider a user wishing to sell a concert ticket. Not only must she find someone who is interested in the concert, but she must also make sure she can trust this person to pay for it. Social Market (SM) solves this problem by allowing users to identify and build connections to other users that can provide interesting goods or information and that are also reachable through a trusted path on an explicit social network like Facebook. This year, we extended the contributions presented in 2011, by introducing two novel distributed protocols that combine interest-based connections between users with explicit links obtained from social networks ala Facebook. Both protocols build trusted multi-hop paths between users in an explicit social network supporting the creation of semantic overlays backed up by social trust. The first protocol, TAPS2, extends our previous work on *TAPS* (Trust-Aware Peer Sampling), by improving the ability to locate trusted nodes. Yet, it remains vulnerable to attackers wishing to learn about trust values between arbitrary pairs of users. The second protocol, PTAPS (*Private TAPS*), improves TAPS2 with provable privacy guarantees by preventing users from revealing their friendship links to users that are more than two hops away in the social network. In addition to proving this privacy property, we evaluate the performance of our protocols through event-based simulations, showing significant improvements over the state of the art. In addition to our previous publication on this topic, our recent work led to a paper that appeared in TCS [20].

6.2.4. Privacy-preserving P2P collaborative filtering

Participants: Davide Frey, Anne-Marie Kermarrec, Antoine Rault, François Taïani.

The huge amount of information available at any time in our connected society calls for a mechanism to filter it efficiently. Recommendation systems provide such a mechanism by personalizing the information displayed for each user. However, the collection of personal information by recommendation systems threatens the privacy of users. We address the two needs for recommendation and privacy through a peer-to-peer user-based collaborative filtering system. Recommendation is done ala GOSSPLE by building an overlay network which connects users with similar interests via clustering and random peer sampling. This overlay network is then used to make recommendations based on what similar users liked. Users' privacy is protected in two ways. Users are protected from a Big Brother adversary by the peer-to-peer design of the system in which profiles are stored only by their owners. Users are protected from other malicious users who would try to learn the content of their profiles by our landmark-based cosine similarity measure. It indirectly computes the similarity of two users by comparing their respective similarities with a set of randomly generated profiles, called landmarks. Thus, users can compute their similarity without revealing their profile, contrarily to the regular cosine similarity when used in a peer-to-peer system.

6.2.5. Gossip protocols for renaming and sorting

Participants: George Giakkoupis, Anne-Marie Kermarrec.

In [33] we devise efficient gossip-based protocols for some fundamental distributed tasks. The protocols assume an n-node network supporting point-to-point communication, and in every round, each node exchanges information of size $O(\log n)$ bits with (at most) one other node. We first consider the *renaming* problem, that is, to assign distinct IDs from a small ID space to all nodes of the network. We propose a renaming protocol that divides the ID space among nodes using a natural push or pull approach, achieving logarithmic round complexity with ID space $\{1, \cdots, (1+\epsilon)n\}$, for any fixed $\epsilon > 0$. A variant of this protocol solves the *tight* renaming problem, where each node obtains a unique ID in $\{1, \cdots, n\}$, in $O(\log^2 n)$ rounds. Next we study the following *sorting* problem. Nodes have consecutive IDs 1 up to n, and they receive numerical values as inputs. They then have to exchange those inputs so that in the end the input of rank k is located at the node with ID k. Jelasity and Kermarrec (2006) suggested a simple and natural protocol, where nodes exchange values with peers chosen uniformly at random, but it is not hard to see that this protocol requires $\Omega(n)$ rounds. We prove that the same protocol works in $O(\log^2 n)$ rounds if peers are chosen according to a non-uniform power law distribution.

This work has been done in collaboration with Philipp Woelfel.

6.2.6. Adaptive streaming

Participants: Ali Gouta, Anne-Marie Kermarrec.

HTTP Adaptive Streaming (HAS) is gradually being adopted by Over The Top (OTT) content providers. In HAS, a wide range of video bitrates of the same video content are made available over the internet so that clients' players pick the video bitrate that best fit their bandwidth. Yet, this affects the performance of some major components of the video delivery chain, namely CDNs or transparent caches since several versions of the same content compete to be cached. We investigated the benefits of a Cache Friendly HAS system (CF-DASH), which aims to improve the caching efficiency in mobile networks and to sustain the quality of experience of mobile clients. We presented a set of observations we made on large number of clients requesting HAS contents [34], [35]. Then, we evaluated CF-dash based on trace-driven simulations and testbed experiments. Our validation results are promising. Simulations on real HAS traffic show that we achieve a significant gain in hit-ratio that ranges from 15% up to 50%.

Work was done in collaboration with Yannick Le Louedec, Zied Aouini and Diallo Mamadou.

6.2.7. DynaSoRe: Efficient in-memory store for social applications

Participant: Arnaud Jégou.

Social network applications are inherently interactive, creating a requirement for processing user requests fast. To enable fast responses to user requests, social network applications typically rely on large banks of cache servers to hold and serve most of their content from the cache. The objective of this work is to build a memory cache system for social network applications that optimizes data locality while placing user views across the system. We call this system DynaSoRe (Dynamic Social stoRe). DynaSoRe storage servers monitor access traffic and bring data frequently accessed together closer in the system to reduce the processing load across cache servers and network devices. Our simulation results considering realistic data center topologies show that DynaSoRe is able to adapt to traffic changes, increase data locality, and balance the load across the system. The traffic handled by the top tier of the network connecting servers drops by 94% compared to a static assignment of views to cache servers while requiring only 30% additional memory capacity compared to the whole volume of cached data.

This work was conducted in collaboration with Xiao Bai, Flavio Junqueira, and Vincent Leroy. The product of this collaboration led to the publication of a paper at the Middleware 2013 conference [26].

6.2.8. Adaptive metrics on distributed recommendation systems

Participants: Anne-Marie Kermarrec, François Taïani, Juan Manuel Tirado Martin.

Current distributed recommendation systems are metric based. This means that recommendation quality depends on a single user comparison function. This is a simple solution that cannot cover the particularities of each system. Classically computing intensive data-mining methods have been used in the field of recommendation. However, they are not proper in distributed scenarios due to the lack of a global vision and the existing restrictions in terms of computing power. In this project, we study how to provide and model ad-hoc similarity metrics that can be automatically adapted to a different number of scenarios. We study our solution from two different points of view: recommendation and performance. In the first, we evaluate the capacity of data mining technics to give users relevant recommendations. Second, by exploring the performance of different approaches in order to obtain relevant recommendations we plan to study the trade-off between relevant recommendations and computational cost.

6.2.9. Cliff-Edge Consensus: Agreeing on the precipice

Participants: Michel Raynal, François Taïani.

In this project, we worked on a new form of consensus that allows nodes to agree locally on the extent of crashed regions in networks of arbitrary size. One key property of our algorithm is that it shows local complexity, i.e. its cost is independent of the size of the complete system, and only depends on the shape and extent of the crashed region to be agreed upon. In [40], we motivate the need for such an algorithm, formally define this new consensus problem, propose a fault-tolerant solution, and prove its correctness.

This work was done in collaboration with Geoff Coulson and Barry Porter.

6.2.10. Clustered network coding

Participants: Fabien André, Anne-Marie Kermarrec, Konstantinos Kloudas, Alexandre Van Kempen.

Modern storage systems now typically combine plain replication and erasure codes to reliably store large amount of data in datacenters. Plain replication allows a fast access to popular data, while erasure codes, e.g. Reed-Solomon codes, provide a storage-efficient alternative for archiving less popular data. Although erasure codes are now increasingly employed in real systems, they experience high overhead during maintenance, i.e. upon failures, typically requiring files to be decoded before being encoded again to repair the encoded blocks stored at the faulty node.

In this work, we propose a novel erasure code system, tailored for networked archival systems. The efficiency of our approach relies on a combination of the use of random codes coupled with a clever yet simple clustered placement strategy. Our repair protocol leverages network coding techniques to reduce by 50% the amount of data transferred during maintenance, as several cluster files are repaired simultaneously. We demonstrate both through an analysis and extensive experimental study conducted on a public testbed that our approach dramatically decreases both the bandwidth overhead during the maintenance process and the time to repair data lost upon failure.

This has been done in collaboration with Erwan le Merrer, Nicolas, Le Scouarnec and Gilles Straub.

ASCOLA Project-Team

6. New Results

6.1. Software composition

Participants: Akram Ajouli, Diana Allam, Ronan-Alexandre Cherrueau, Rémi Douence, Hervé Grall, Florent Marchand de Kerchove de Denterghem, Jacques Noyé, Jean-Claude Royer, Mario Südholt.

6.1.1. Service-oriented computing

Services are frequently implemented using object-oriented frameworks. In this context, two properties are particularly important: (i) a loose coupling between the service layer and the object layer, allowing evolution of the service layer with a minimal impact on the object layer, (ii) interoperability induced by the substitution principle associated to subtyping in the object layer, thus allowing to freely convert a value of a subtype into a supertype. However, through experimentation with Apache's popular service framework CXF, we observed some undesirable coupling and interoperability issues due to the failure of the substitution principle [23]. Therefore we have proposed a new specification method for the data binding used to translate data between the object and service layers [24]. We have shown that if the CXF framework follows the specification, the substitution principle is satisfied, with all its advantages.

6.1.2. Modularity and program transformations

Refactoring tools are commonly used for remodularization tasks. Basic refactoring operations are combined to perform complex program transformations, but the resulting composed operations are rarely reused, even partially, because popular tools have few support for composition. In [31], we have recast two calculus for static composition of refactorings in a type system framework and we have discussed their use for inferring useful properties. We have illustrated the value of support for static composition in refactoring tools with a complex remodularization use case: a round-trip transformation between programs conforming to the Composite and Visitor patterns. Composite and Visitor design patterns have dual properties with respect to modularity, thus they are good candidates to explore their transformations. In [22] we have extended our initial refactoring-based round-trip transformation between these two structures and we have studied how that transformation is impacted by four variations in the implementation of these patterns. We have validated that study by computing the smallest preconditions for the resulting transformations. We have also automated the transformation and applied it to JHotDraw, where the studied variations occur. Finally, [11] presents more exhaustively modular transformations and design patterns. We have also proposed a reversible transformation in the Singleton pattern to benefit from optimization by introducing this pattern and flexibility by its suppression according to the requirements of the software user.

6.1.3. Domain specific languages

In the context of Charles Prud'hommes' PhD Thesis, we have developed a domain specific language in order to specify strategies of filtering propagation in constraint solvers. Indeed, constraint programming replaces brute force generate-and-test by the exploration of the solution space based on incremental instantiation and constraint propagation. Strategies of incremental instantiation (also known as heuristics) have been heavily studied. However, most solvers propagate constraints with a simple fix point computation based on a queue of constraints to propagate (or several queues in order to deal with the grain/cost of filtering algorithms). This technique has a good behavior in general but for a given problem a dedicated strategy can be more efficient. Our declarative DSL and its support in the new version of the constraint solver Choco [19], [52] enables us to easily experiment with different propagation strategies. Moreover, our DSL supports properties such as completeness, intended incompleteness or non ambiguity.

6.1.4. Constructive security

In the field of techniques for the development of secure software systems we have presented results on the enforcement of security properties in service-oriented systems and Javascript programs.

Concerning the security of service-based systems, we have first presented a software framework that harnesses a type based policy language and aspect-based support for protocol adaptation in service-oriented systems by means of flexible reference monitors [29], [28]. We have shown how this framework improves the security, interoperability and evolution issues of service systems using the OAuth 2.0 standard for the authorization of resource accesses. The OAuth 2 protocol is a recent IETF standard devoted to providing authorization to clients requiring access to specific resources over HTTP. It was recently adopted by major internet companies and software editors, such as Google, Facebook, Microsoft, and SAP. We have shown how to improve the security of software systems that use OAuth 2 in the presence of different kinds of clients.

Furthermore, we have developed a new notion of transformation operators, so-called workflow adaptation schemas (WASs) for service compositions that facilitates the integration and modification of security functionalities of service-oriented systems [30]. These schemas may be generic and specialized through parameter instantiation. A set of schemas therefore effectively provides a domain-specific language for the transformation of service-oriented applications. We have developed a set of specific schemas and applied them to the OAuth 2 standard in order to implement state-based security hardening strategies. We have also implemented tool support for WASs and implemented some of the security scenarios involving OAuth 2 (see Sec. 5.4).

Finally, we have shown that a wide range of strategies to make secure JavaScript-based applications can be described pertinently using aspects [42]. To this end, we have reviewed major categories of approaches to make client-side applications secure and have discussed uses of aspects that exist for some of them. We also propose aspect-based techniques for the categories that have not been studied previously. We have given examples of applications where aspects are useful as a general means to flexibly express and implement security policies for JavaScript.

6.2. Aspect-Oriented Programming

Participants: Rémi Douence, Ismael Figueroa, Jacques Noyé, Mario Südholt, Nicolas Tabareau, Jurgen Van Ham.

6.2.1. Aspects in a concurrent and distributed setting

Aspect oriented programming modularizes crosscutting concerns by gathering several join points. In the context of distributed applications these point cuts can be on different machines. In this case, a sequence of join points must be defined as a sequence of logical joint points (à la Lamport). We propose an aspect oriented languages to define distributed aspects in JavaScript in a distributed context. Our proposal [18] is based on vector clocks in order to logically relate join points and can ignore "illogical" (that is late or early) join points. In can also enforce causal communications when no join point must be discarded. We have exemplified the advantages of our technique with different applications such as a discussion forum, a retweet scenario and a web browser.

Multiparty session types allow the definition of distributed processes with strong communication safety properties. A global type is a choreographic specification of the interactions between peers, which is then projected locally in each peer. Well-typed processes behave accordingly to the global protocol specification. Multiparty session types are however monolithic entities that are not amenable to modular extensions. Also, session types impose conservative requirements to prevent any race condition, which prohibit the uniform application of extensions at different points in a protocol. We have proposed a means to support modular extensions with *aspectual session types* [47], a static pointcut/advice mechanism at the session type level. To support the modular definition of crosscutting concerns, we augment the expressivity of session types to allow harmless race conditions. We formally prove that well-formed aspectual session types entail communication safety. As a result, aspectual session types make multiparty session types more flexible, modular, and extensible.

We have added dedicated concurrency support to EScala, our extension of Scala that introduces composable *declarative events* as a way to integrate Aspect-Oriented Programming and Event-Based Programming in the context of Object-Oriented Programming. In JEScala, Events, which were synchronous in EScala, can be declared as *asynchronous* so that they are handled concurrently to their emitter. Moreover, two new operators, a join and a choice operator, inherited from the join calculus - hence the name of the new prototype, can now be used to compose events and control concurrency. In [48], we present JEScala, show that it captures coordination schemas in a more expressive and modular way than plain join languages and provide a first performance assessment.

6.2.2. Effective aspects

We have proposed a novel approach to embed pointcut/advice aspects in a typed functional programming language like Haskell. Aspects are first-class, can be deployed dynamically, and the pointcut language is extensible. Type soundness is guaranteed by exploiting the underlying type system, in particular phantom types and a new anti-unification type class. The use of monads brings type-based reasoning about effects for the first time in the pointcut/advice setting and enables modular extensions of the aspect language [46], [16].

To allow a type-safe embedding of aspects in Haskell, we had to develop a notion of anti-unification in Haskell type system. The anti-unification problem is that of finding the most specific pattern of two terms. While dual to the unification problem, anti-unification has rarely been considered at the level of types. We have developed an algorithm to compute the least general type of two types in Haskell, using the logic programming power of type classes [53]. That is, we have defined a type class for which the type class instances resolution performs anti-unification.

6.2.3. Reasoning about aspect interference

When a software system is developed using several aspects, special care must be taken to ensure that the resulting behavior is correct. This is known as the *aspect interference problem*, and existing approaches essentially aim to detect whether a system exhibits problematic interferences of aspects. We have described how to control aspect interference by construction by relying on the type system. More precisely, we combine a monadic embedding of the pointcut/advice model in Haskell with the notion of membranes for aspect-oriented programming [34]. Aspects must explicitly declare the side effects and the context they can act upon. Allowed patterns of control flow interference are declared at the membrane level and statically enforced. Finally, computational interference between aspects is controlled by the membrane topology. To combine independent and reusable aspects and monadic components into a program specification we use *monad views*, a recent technique for conveniently handling the monadic stack.

Oliveira and colleagues recently developed a powerful model to reason about mixin-based composition of effectful components and their interference, exploiting a wide variety of techniques such as equational reasoning, parametricity, and algebraic laws about monadic effects. Our work addresses the issue of reasoning about interference with effectful aspects in the presence of unrestricted quantification through pointcuts. While global reasoning is required, we have shown that it is possible to reason in a compositional manner, which is key for the scalability of the approach in the face of large and evolving systems. We have established a general equivalence theorem that is based on a few conditions that can be established, reused, and adapted separately as the system evolves. Interestingly, one of these conditions, local harmlessness, can be proven by a translation to the mixin setting, making it possible to directly exploit previously established results about certain kinds of harmless extensions [33].

In aspect-oriented programming (AOP) languages, advice evaluation is usually considered as part of the base program evaluation. While viewing aspects as part of base level computation clearly distinguishes AOP from reflection, it also comes at a price: because aspects observe base level computation, evaluating pointcuts and advice at the base level can trigger infinite regression. To avoid these pitfalls, we have introduced levels of execution in the programming language, thereby allowing aspects to observe and run at specific, possibly different, levels. We adopt a defensive default that avoids infinite regression, and gives advanced programmers the means to override this default using level-shifting operators [21].

6.3. Resource management in Cloud computing

Participants: Frederico Alvares, Gustavo Bervian Brand, Yousri Kouki, Adrien Lèbre, Thomas Ledoux, Guillaume Le Louët, Jean-Marc Menaud, Jonathan Pastor, Flavien Quesnel, Mario Südholt.

We have contributed on several topics: multiple autonomic managers for Cloud infrastructure, SLA management for Cloud elasticity, fully distributed and autonomous virtual machine scheduling, and simulator toolkits for IaaS platforms.

6.3.1. Cloud infrastructure based on multiple autonomic managers

One of the main reasons for the wide adoption of Cloud Computing is the concept of elasticity. Implementing elasticity to tackle varying workloads while optimizing infrastructures (e.g. utilization rate) and fulfilling the application requirements on Quality of Service should be addressed by self-adaptation techniques able to manage complexity and dynamism. However, since Cloud systems are organized in different but dependent Cloud layers, self-management decisions taken in isolation in a certain layer may indirectly interfere with the decision taken by an other layer. Indeed, non-coordinated managers may lead to conflicting decisions and consequently to non-desired states.

We have proposed a framework for the coordination of multiple autonomic managers in cloud environments [25]. The PhD thesis of Frederico Alvares [12], defended in April 2013, is based on this framework. This thesis proposes a self-adaptation approach that considers both application internals (architectural elasticity) and infrastructure (resource elasticity), managed by multiple autonomic managers, to reduce the energy footprint in Cloud infrastructures.

6.3.2. SLA Management for Cloud elasticity

Elasticity is the intrinsic element that differentiates Cloud Computing from traditional computing paradigms, since it allows service providers to rapidly adjust their needs for resources to absorb the demand and hence guarantee a minimum level of Quality of Service (QoS) that respects the Service Level Agreements (SLAs) previously defined with their clients. However, due to non-negligible resource initiation time, network fluctuations or unpredictable workload, it becomes hard to guarantee QoS levels and SLA violations may occur. The main challenge of service providers is to maintain its consumer's satisfaction while minimizing the service costs due to resources fees. The PhD thesis of Yousri Kouki [13], defended in December, proposes different contributions to address this issue: CSLA, a specific language to describe SLA for Cloud services ; HybridScale, an auto-scaling framework driven by SLA [39], [17].

6.3.3. Fully Distributed and Autonomous Virtualized Environments

We have consolidated the DVMS system to obtain a fully distributed virtual machine scheduler [44]. This system makes it possible to schedule VMs cooperatively and dynamically in large scale distributed systems. Simulations (up to 64K VMs) and real experiments both conducted on the Grid'5000 large-scale distributed system [44] showed that DVMS is scalable. This building block is a first element of a more complete cloud OS, entitled DISCOVERY (DIStributed and COoperative mechanisms to manage Virtual EnviRonments autonomicallY) [56]. The ultimate goal of this system is to overcome the main limitations of the traditional server-centric solutions. The system, currently under investigation in the context of the Jonathan Pastor's PhD, relies on a peer-to-peer model where each agent can efficiently deploy, dynamically schedule and periodically checkpoint the virtual environments it manages.

6.3.4. Testing the cloud

Computer science, as other sciences, needs instruments to validate theoretical research results, as well as software developments. Although simulation and emulation are generally used to get a glance of the behavior of new algorithms, they use over-simplified models in order to reduce their execution time and thus cannot be accurate enough. Leveraging a scientific instrument to perform actual experiments is an undeniable advantage. However, conducting experiments on real environments is still too often a challenge for researchers, students, and practitioners: first, because of the unavailability of dedicated resources, and second, because of the inability to create controlled experimental conditions, and to deal with the wide variability of software

requirements. During 2013, we have contributed to a new topic addressing the "testing the cloud" challenge. First, we have presented the latest mechanisms we have designed to enable the automated deployment of the major open-source IaaS cloudkits (i.e., Nimbus, OpenNebula, CloudStack, and OpenStack) on Grid'5000 [26]. Providing automatic, isolated and reproducible deployments of cloud environments lets end-users study and compare each solution or simply leverage one of them to perform higher-level cloud experiments (such as investigating Map/Reduce frameworks or applications). Moreover, we have presented EXECO, a library that provides easy and efficient control of large scale experiments through a set of tools well as tools designed for scripting distributed computing experiments on any computing platform. We have illustrated its interest by presenting two experiments dealing with virtualization technologies on the Grid'5000 testbed [37].

6.3.5. Adding virtualization abstractions into the Simgrid toolkit

In the context of the ANR SONGS project and in collaboration with Takahiro Hirofuchi, researcher at AIST (Japan), we have extended the Simgrid framework to be able to simulate virtualized distributed infrastructures [35]. In addition, we have proposed the first class support of live migration operations within such a simulator toolkit for large scale distributed infrastructures. We have developed a resource share calculation mechanism for VMs and a live migration model implementing the precopy migration algorithm of Qemu/KVM. We have confirmed that our simulation framework correctly reproduced live migration behaviors of the real world under various conditions [36].

6.3.6. Power and energy management in the cloud

Power management has become one of the main challenges for data center infrastructures. Currently, the cost of powering a server is approaching the cost of the server hardware itself, and, in a near future, the former will continue to increase, while the latter will go down. In this context, virtualization is used to decrease the number of servers, and increase the efficiency of the remaining ones.

First, in [43] we have proposed an approach and a model to estimate the total power consumption of a virtual machine, by taking into account its static (e.g. memory) and dynamic (e.g. CPU) consumption of resources. Second, we have rewritten the Entropy framework (in OptiPlace) to give it the support of external models, named views. Entropy, based on the Constraint Programming solver Choco written in Java, does not really scale well. We have studied Entropy's scalability properties [32] and have then integrated heuristics and constraints in OptiPlace [40].

The evaluation of these policies on real infrastructures has become an important and difficult issue. The corresponding techniques have become so complex that there is a need for load injection frameworks able to inject resource load in a tested datacenter instead of model-driven simulation. For this reason we have developed StressCloud [41], [51], a framework to manipulate the activities of a group of Virtual Machines and observe the resulting performance.

ASPI Project-Team

5. New Results

5.1. Iterative isotone regression

Participant: Arnaud Guyader.

This is a collaboration with Nicolas Hengartner (Los Alamos), Nicolas Jégou (université de Rennes 2) and Eric Matzner–Løber (université de Rennes 2), and with Alexander B. Németh (Babeş Bolyai University) and Sándor Z. Németh (University of Birmingham).

We explore some theoretical aspects of a recent nonparametric method for estimating a univariate regression function of bounded variation. The method exploits the Jordan decomposition which states that a function of bounded variation can be decomposed as the sum of a non-decreasing function and a non-increasing function. This suggests combining the backfitting algorithm for estimating additive functions with isotonic regression for estimating monotone functions. The resulting iterative algorithm is called IIR (iterative isotonic regression). The main result in this work [22] states that the estimator is consistent if the number of iterations k_n grows appropriately with the sample size n. The proof requires two auxiliary results that are of interest in and by themselves: firstly, we generalize the well-known consistency property of isotonic regression to the framework of a non-monotone regression function, and secondly, we relate the backfitting algorithm to the von Neumann algorithm in convex analysis. We also analyse how the algorithm can be stopped in practice using a data-splitting procedure.

With the geometrical interpretation linking this iterative method with the von Neumann algorithm, and making a connection with the general property of isotonicity of projection onto convex cones, we derive in [14] another equivalent algorithm and go further in the analysis.

5.2. Mutual nearest neighbors

Participant: Arnaud Guyader.

This is a collaboration with Nicolas Hengartner (Los Alamos).

Motivated by promising experimental results, this work [13] investigates the theoretical properties of a recently proposed nonparametric estimator, called the MNR (mutual nearest neighbors) rule, which estimates the regression function $m(x) = E[Y|X = x]$ as follows: first identify the k nearest neighbors of x in the sample, then keep only those for which x is itself one of the k nearest neighbors, and finally take the average over the corresponding response variables. We prove that this estimator is consistent and that its rate of convergence is optimal. Since the estimate with the optimal rate of convergence depends on the unknown distribution of the observations, we also have adaptation results by data-splitting.

5.3. Adaptive multilevel splitting

Participants: Frédéric Cérou, Arnaud Guyader, Florent Malrieu.

This is a collaboration with Pierre Del Moral (EPI ALEA, Inria Bordeaux—Sud Ouest).

We show that an adaptive version of multilevel splitting for rare events is strongly consistent. We also show that the estimates satisfy a CLT (central limit theorem), with the same asymptotic variance as the non-adaptive algorithm with the optimal choice of the parameters. It is a strong and general result, that generalizes some of our previous results, and the proof is quite technical and involved.

5.4. Total variation estimates for the TCP process

Participant: Florent Malrieu.

This is a collaboration with Jean-Baptiste Bardet (université de Rouen), Alejandra Christen (University of Chile), Arnaud Guillin (université de Clermont–Ferrand), and Pierre–André Zitt (université de Paris–Est Marne–la–Vallée).

The TCP window size process appears in the modeling of the famous Transmission Control Protocol used for data transmission over the Internet. This continuous time Markov process takes its values in $[0, \infty)$, is ergodic and irreversible. The sample paths are piecewise linear deterministic and the whole randomness of the dynamics comes from the jump mechanism. The aim of [27] is to provide quantitative estimates for the exponential convergence to equilibrium, in terms of the total variation and Wasserstein distances, using coupling methods. The technique could be applied to a large class of Markov processes as well.

5.5. On the stability of planar randomly switched systems

Participant: Florent Malrieu.

This is a collaboration with Michel Benaïm (université de Neuchâtel), Stéphane Le Borgne (IRMAR) and Pierre–André Zitt (université de Paris–Est Marne–la–Vallée).

The paper [28] illustrates some surprising instability properties that may occur when stable ODE's are switched using Markov dependent coefficients. Consider the random process (X_t) solution of $dX_t/dt = A(I_t)X_t$ where (I_t) is a Markov process on $\{0, 1\}$ and A_0 and A_1 are real Hurwitz matrices on $\mathbb{R}^2$. Assuming that there exists $\lambda \in (0, 1)$ such that $(1 - \lambda)A_0 + \lambda A_1$ has a positive eigenvalue, we establish that the norm of X_t may converge to 0 or infinity, depending on the the jump rate of the process I. An application to product of random matrices is studied. This work can be viewed as a probabilistic counterpart of the paper [26] by Baldé, Boscain and Mason.

5.6. Marginalization in rare event simulation for switching diffusions

Participant: François Le Gland.

This is a collaboration with Anindya Goswami (IISER, Poone).

Switching diffusions are continuous–time Markov processes with a hybrid continuous / finite state space. A rare but critical event (such as a scalar function of the continuous component of the state exceeding a given threshold) can occur for several reasons:

- the process can remain in *nominal* mode, where the critical event is very unlikely to occur,
- or the process can switch in some *degraded* mode, where the critical event is much more likely to occur, but the switching itself is very unlikely to occur.

Not only is it important to accurately estimate the (very small) probability that the critical event occurs before some fixed final time, but it is also important to have an accurate account on the reason why it occured, or in other words to estimate the probability of the different modes. A classical implementation of the multilevel splitting would not be efficient. Indeed, as soon as (even a few) samples paths switch to a *degraded* mode, these sample paths will dominate and it will not be possible to estimate the contribution of samples paths in the *nominal* mode. Moreover, sampling the finite component of the state is not efficient to accurately estimate the (very small) probability of rare but critical modes. A more efficient implementation is based on marginalization, i.e. in sampling jointly the continuous component and the probability distribution of the finite component given the past continuous component [18]. The latter is a probability vector, known as the Wonham filter, that satisfies a deterministic equation.

5.7. Combining importance sampling and multilevel splitting for rare event simulation

Participants: François Le Gland, Damien–Barthélémy Jacquemart.

This is a collaboration with Jérôme Morio (ONERA, Palaiseau).

The problem is to accurately estimate the (very small) probability that a rare but critical event (such as a scalar function of the state exceeding a given threshold) occurs before some fixed final time. Multilevel splitting is a very efficient solution, in which sample paths are propagated and are replicated when some intermediate events occur. Events that are defined in terms of the state variable only (such as a scalar function of the state exceeding an intermediate threshold) are not a good design. A more efficient but more complicated design would be to let the intermediate events depend also on time. An alternative design is to keep intermediate events simple, defined in terms of the state variable only, and to make sure that samples that exceed the threshold early are replicated more than samples that exceed the same threshold later [19].

5.8. Sequential data assimilation: ensemble Kalman filter vs. particle filter

Participants: François Le Gland, Valérie Monbet.

The contribution has been to prove (by induction) the asymptotic normality of the estimation error, i.e. to prove a central limit theorem for the ensemble Kalman filter. Explicit expression of the asymptotic variance has been obtained for linear Gaussian systems (where the exact solution is known, and where EnKF is unbiased). This expression has been compared with explicit expressions of the asymptotic variance for two popular particle filters: the bootstrap particle filter and the so–called optimal particle filter, that uses the next observation in the importance distribution.

5.9. Non–homogeneous Markov–switching models

Participant: Valérie Monbet.

This is a collaboration with Pierre Ailliot (université de Bretagne occidentale, Brest).

We have developped various hidden non–homogeneous Markov–switching models for description and simulation of univariate and multivariate time series. Considered application are in weather variables modelling but also in economy. The main originality of the proposed models is that the hidden Markov chain is not homogeneous, its evolution depending on the past wind conditions or other covariates. It is shown that it permits to reproduce complex non–linearities.

5.10. Dynamical partitioning of directional ocean wave spectra

Participant: Valérie Monbet.

This is a collaboration with Pierre Ailliot (université de Bretagne occidentale, Brest) and Christophe Maisondieu (IFREMER, Brest).

Directional wave spectra generally exhibit several peaks due to the coexistence of wind sea generated by local wind conditions and swells originating from distant weather systems. The paper [24] proposes a new algorithm for partitioning such spectra and retrieving the various systems which compose a complex sea-state. It is based on a sequential Monte Carlo algorithm which allows to follow the time evolution of the various systems. The proposed methodology is validated on both synthetic and real spectra and the results are compared with a method commonly used in the literature.

5.11. Track–before–detect

Participants: François Le Gland, Alexandre Lepoutre.

This is a collaboration with Olivier Rabaste (ONERA, Palaiseau).

The problem considered in [20] is tracking one or several targets in a track–before–detect (TBD) context using particle filters. These filters require the computation of the likelihood of the complex measurement given the target states. This likelihood depends on the complex amplitudes of the targets. When the complex amplitude fluctuates over time, time coherence of the target cannot be taken into account. However, for the single target case, spatial coherence of this amplitude can be taken into account to improve the filter performance, by marginalizing the likelihood of the complex measurement over the amplitude parameter. The marginalization depends on the fluctuation law considered. We show that for the Swerling 1 model the likelihood of the complex measurement can be obtained analytically in the multi-target case. For the Swerling 0 model no closed form can be obtained in the general multi–target setting. Therefore we resort to some approximations to solve the problem. Finally, we demonstrate with Monte Carlo simulations the gain of this method both in detection and in estimation compared to the classic method that works with the square modulus of the complex signal.

The problem considered in [21] is detecting and tracking a single radar target with amplitude fluctuation Swerling 1 and 3 in a track–before–detect context with particle filter. Those fluctuations are difficult to take into account as they are uncoherent from measurement to measurement. Thus, conventionnal filters work on square modulus of the complexe signal to remove the unknown phase of complex amplitude and the marginalized over the law of the modulus but they lose the spatial coherence of the amplitude in the measurement. We show in this paper that complex measurements can be marginalized directly while taking into account the spatial coherence of the complex amplitude. Finally, we show the benefit of this method both in detection and in estimation via Monte Carlo simulations.

ATLANMOD Project-Team

6. New Results

6.1. Reverse Engineering

Model Driven Reverse Engineering (MDRE), and its applications such as software modernization, is a discipline in which model-driven development (MDD) techniques are used to treat legacy systems. During this year, Atlanmod has continued working actively on this research area. The main contributions are the following:

- In the context of the ARTIST FP7 project, the work has started on reusing (and extending accordingly) MoDisco and several of its components to provide the Reverse Engineering support required within the project. More particularly, the MoDisco Model Discovery + Model Understanding two-step approach is being promoted as an important part of the ARTIST migration methodology and process [35] [19]. Work has also been performed, in the context of the TEAP FUI project dealing with Enterprise Architecture, on how to design and implement a model driven federation approach from heterogeneous data sources (e.g. Excel files, databases, etc.) directly inspiring from these same MoDisco principles [20].
- In order to react to the ever-changing market, every organization needs to periodically reevaluate and evolve its company policies. These policies must be enforced by its Information System (IS) by means of a set of so-called business rules that drive the system behavior and data. Clearly, policies and rules must be aligned at all times but unfortunately this is a challenging task. In most ISs, the implementation of business rules is scattered among the different components of the system, therefore appropriate techniques must be provided for the discovery and evolution of changing business rules. In [39], [25], [26], we describe a MDRE framework and tool aiming at extracting business rules out of COBOL source code. In [27], we describe a Model-based process and tool to extract business rules, expressed as OCL integrity constraints, from relational databases. In these works, the use of modeling techniques facilitate the representation of the rules at a higher-abstraction level which enables stakeholders to understand and manipulate them more easily. A thesis financed by IBM to advance the research on this topic has been completed this year
- In a web context, JSON has become a very popular lightweigth format for data exchange. JSON is human readable and easy for computers to parse and use. However, JSON is schemaless. Though this brings some benefits (e.g. flexibility in the representation of the data) it can become a problem when consuming and integrating data from different JSON services since developers need to be aware of the structure of the schemaless data. We believe that a mechanism to discover (and visualize) the implicit schema of the JSON data would largely facilitate the creation and usage of JSON services. For instance, this would help developers to understand the links between a set of services belonging to the same domain or API. In this sense, we have proposed a model-based approach to generate the underlying schema of a set of JSON documents [22].

6.2. Security

Most companies information systems are composed by heterogeneous components responsible of hosting, creating or manipulating critical information for the day-to-day operation of the company. Securing this information is therefore one of their main concerns, more particularly specifying Access Control (AC) policies. However, the task of implementing an AC security policy (sometimes relying on several mechanisms) remains complex and error prone as it requires knowing low level and vendor-specific facilities. In this context, discovering and understanding which security policies are actually being enforced by the Information System (IS) becomes critical. Thus, the main challenge consists in bridging the gap between the vendor-dependent security features and a higher-level representation. This representation has to express the policies by abstracting from the specificities of the system components, allowing security experts to better understand the policy and to implement all related evolution, refactoring and manipulation operations in a reusable way.

In 2013, we have tackled the aforementioned problems with respect to three key information system components: networks of firewalls, relational database systems and content management systems.

- Firewalls are a key element in network security. They are in charge of filtering the traffic of the network in compliance with a number of access-control rules that enforce a given security policy. In [33] we have described a model-driven reverse engineering approach able to extract the security policy implemented by a set of firewalls in a working network, easing the understanding, analysis and evolution of network security policies. In [17] we have extended this method to cope with a more complex and specific scenario, i.e, the management of stateful packet filtering.
- A similar approach have been successfully used to extract AC information from relational database systems. Concretely, in [32] we contribute a security metamodel and a reverse engineering process that combines standard database access-control rules with the fine-grained access control provided by triggers and stored procedures. The extraction of this comprehensive model helps security experts to visualize and manipulate database security policies in a vendor-independent manner.
- Out-of-the-box Web Content Management Systems (WCMSs) are the tool of choice for the development of millions of enterprise web sites. However, little attention has been brought to the analysis of how developers use the content protection mechanisms provided by WCMSs, in particular, Access-control (AC). We have proposed in [34] a metamodel tailored to the representation of WCMS AC policies, easing the analysis and manipulation tasks by abstracting from vendor-specific details.

6.3. Collaborative development

In the field of Domain-Specific Languages (DSLs), we have focused on the improvement of the DSLs definition process. When developing DSMLs, the participation of end-users is normally limited to providing domain knowledge and testing the resulting language prototypes. Language developers, which are perhaps not domain experts, are therefore in control of the language development and evolution. This may cause misinterpretations which hamper the development process and the quality of the DSML. Thus, it would be beneficial to promote a more active participation of end-users in the development process of DSMLs. While current DSML workbenches are mono-user and designed for technical experts, we have presented a process and tool support for the example-driven, collaborative construction of DSMLs based on Collaboro in order to engage end-users in the creation of their own languages [23], [24].

6.4. MDE Scalability

As Model-Driven Engineering (MDE) is increasingly applied to larger and more complex systems, additional research and development is imperative in order to enable MDE to remain relevant with industrial practice. In [31] we attempt to provide a research roadmap for scalability in MDE and outline directions for work in this emerging research area. As a first result in this roadmap, in [37] we show that rule-based languages like ATL have strong parallelization properties. Parallelization is indeed one of the traditional ways of making computation systems scalable. We describe the implementation of a parallel transformation engine for the current version of the ATL language and experimentally evaluate the consequent gain in scalability. Finally in [28] we compare the improved scalability of the ATL transformation engine with other engines in the community by addressing the task of generating and analyzing very large flow graphs.

6.5. Model Quality

Our work aims to enhance the quality of the modeling activity in the context of software engineering and language engineering. This year, this has translated in the following results:

- A benchmark that facilitates the comparison between the plethora of tools that provide some kind of quality assurance for models. Similarly to what it is done in many other domains, a common set of test benchmarks that new tools can rely on to experiment and evaluate themselves could speed up the advance in the field. Our proposal can be found [30]

- Validation of the feasibility to apply this kind of techniques in industrial settings based on two case studies [12] and [36]
- Advanced on the verification of model transformations using SMT solvers (instead of SAT or CSP-based approaches commonly used before), with some encouraging results [21] and, related to this, [13]
- A method to build models using instance-level information in terms of examples and counterexamples (gathering requirements using these instance scenarios is usually better from a stakeholder's point of view than trying to explain us general rules about the business). So far existing approaches have often focused on the generation of static models from such instance-level information but have omitted the inference of OCL business rules that could complement the static models and improve the precision of the software specification. We propose an approach to automating such inference [29]. The basic idea is based on an incorporation of the problem solving mechanism and getting user feedback: Candidates are generated by a problem solving, and irrelevant ones are eliminated using the user feedback on generated counterexamples and examples. Our approach is realized with the support tool InferOCL and has been applied on several user cases, indicating a possibility to apply this solution prototype in practice.

CAIRN Project-Team

6. New Results

6.1. Reconfigurable Architecture Design

6.1.1. Arithmetic Operators for Cryptography and Fault-Tolerance

Participants: Arnaud Tisserand, Emmanuel Casseau, Thomas Chabrier, Karim Bigou, Franck Bucheron, Jérémie Métairie, Nicolas Veyrat-Charvillon, Nicolas Estibals.

Arithmetic Operators for Fast and Secure Cryptography. Scalar recoding is popular to speed up ECC (elliptic curve cryptography) scalar multiplication: non-adjacent form, double-base number system, multi-base number system (MBNS). But fast recoding methods require pre-computations: multiples of base point or off-line conversion. In paper [42] presented at ARITH, we presented a multi-base (e.g. (2,3,5,7)) recoding method for ECC scalar multiplication based on i) a greedy algorithm starting least significant terms first, ii) cheap divisibility tests by multi-base elements and iii) fast exact divisions by multi-base elements. Multi-base terms are obtained on-the-fly using a special recoding unit which operates in parallel to curve-level operations and at very high speed. This ensures that all recoding steps are performed fast enough to schedule the next curve-level operations without interruptions. The proposed method can be fully implemented in hardware without pre-computations. We report FPGA implementation details and very good performance compared to state-of-art results. A specific version of our method allows random recodings of the scalar which can be used as a partial counter-measure against side-channel attacks. The PhD thesis defended by Thomas Chabrier [18] deals with MBNS and other types of arithmetic recodings for ECC scalar multiplication (title: "Arithmetic recodings for ECC cryptoprocessors with protections against side-channel attacks").

In the paper [67], presented at ComPAS, we presented efficient arithmetic operators for divisibility tests and modulo operations for large operands (e.g. 160-600 bits like in cryptographic applications) and by a set of small constants such as $(2^a, 3, 5, 7, 9)$ where $1 \leq a \leq 12$. These operators have been validated and implemented on FPGAs.

In the paper [39] presented at CHES, we described a new RNS modular inversion algorithm based on the extended Euclidean algorithm and the plus-minus trick. In our algorithm, comparisons over large RNS values are replaced by cheap computations modulo 4. Comparisons to an RNS version based on Fermat's little theorem were carried out. Comparisons to a version based on Fermat's little theorem were carried out. The number of elementary modular operations is significantly reduced: a factor 12 to 26 for multiplications and 6 to 21 for additions. Virtex 5 FPGAs implementations show that for a similar area, our plus-minus RNS modular inversion is 6 to 10 times faster. Other implementation results of RNS for ECC cryptosystems have been presented in [75] and [74].

ECC Processor with Protections Against SCA. A dedicated processor for elliptic curve cryptography (ECC) is under development. Functional units for arithmetic operations in GF(2^m) and GF(p) finite fields and 160-600-bit operands have been developed for FPGA implementation. Several protection methods against side channel attacks (SCA) have been studied. The use of some number systems, especially very redundant ones, allows one to change the way some computations are performed and then their effects on side channel traces. This work is done in the PAVOIS project.

Arithmetic Operators for Fault Tolerance. In the ARDyT project, we work on computation algorithms, representations of numbers and hardware implementations of arithmetic operators with integrated fault detection (and/or fault tolerance) capabilities. The target arithmetic operators are: adders, subtracters, multipliers (and variants of multiplications by constants, square, FMA, MAC), division, square-root, approximations of the elementary functions. We study two approaches: residue codes and specific bit-level coding in some redundant number systems for fault detection/tolerance integration at the arithmetic operator/unit level. FPGA prototypes are under development.

6.1.2. Reconfigurable Processor Extensions Generation

Participants: Christophe Wolinski, François Charot.

Most proposed techniques for automatic instruction sets extension usually dissociate pattern selection and instruction scheduling steps. The effects of the selection on the scheduling subsequently produced by the compiler must be predicted. This approach is suitable for specialized instructions having a one-cycle duration because the prediction will be correct in this case. However, for multi-cycle instructions, a selection that does not take scheduling into account is likely to privilege instructions which will be, *a posteriori*, less interesting than others in particular in the case where they can be executed in parallel with the processor core. The originality of our research work is to carry out specialized instructions selection and scheduling in a single optimization step. This complex problem is modeled and solved using constraint programming techniques. This approach allows the features of the extensible processor to be taken into account with a high degree of flexibility. Different architectures models can be envisioned. This can be an extensible processor tightly coupled to a hardware extension having a minimal number of internal registers used to store intermediate results, or a VLIW-oriented extension made up of several processing units working in parallel and controlled by a specialized instruction. These techniques have been implemented in the Gecos source-to-source framework.

Novel techniques addressing the interactions between code transformation (especially loops) and instruction set extension are under study. The idea is to automatically transform the original loop nests of a program (using the polyhedral model) to select specialized and vector instructions. These new instructions may use local memories located in the hardware extension and used to store intermediates data produced at a given loop iteration. Such transformations lead to patterns whose effect is to significantly reduce the pressure on the memory of the processor. An experiment realized on the matrix multiplication (extracted from PolyBench/C, the polyhedral benchmark suite) using an Xtensa extensible and configurable processor from Tensilica shows interesting speedups. Speedup of 4.3 for the transformed code compared to the initial code for matrices of size 512x512 and speedup of 8.75 (respectively 20.15) in case of an extension allowing SIMD vector operations on vector of 4 32-bit words (respectively 16 32-bit words) are observed.

6.1.3. Runtime Mapping of Hardware Accelerators on the FlexTiles 3D Self-Adaptive Heterogeneous Manycore

Participants: Olivier Sentieys, Antoine Courtay, Christophe Huriaux.

FlexTiles is a 3D stacked chip with a manycore layer and a reconfigurable layer. This heterogeneity brings a high level of flexibility in adapting the architecture to the targeted application domain for performance and energy efficiency. A virtualisation layer on top of a kernel hides the heterogeneity and the complexity of the manycore and fine-tunes the mapping of an application at runtime. The virtualisation layer provides self-adaptation capabilities by dynamically relocation of application tasks to software on the manycore or to hardware on the reconfigurable area. This self-adaptation is used to optimize load balancing, power consumption, hot spots and resilience to faulty modules. The reconfigurable technology is based on a Virtual Bit-Stream (VBS) that allows dynamic relocation of accelerators just as software based on virtual binary code allows task relocation.

We have proposed a novel approach to hardware task relocation in an FPGA-based reconfigurable fabric, allowing offline design, routing, and unfinalized placement of hardware IPs and dynamic placement of the corresponding bit-streams at run-time. Our proposal relies on a custom dual-context FPGA configuration memory organization in a shift-register manner and on a dedicated bit-stream insertion controller leading to a break-through in terms of adaptive capabilities of the reconfigurable hardware. We show that using our custom shift-register organization across the configuration memory, and under some weak constraints, can greatly reduce the overhead implied by the 1-D to 2-D mapping of the shift-register onto the logic fabric. The use of partial dynamic reconfiguration in FPGA-based systems has grown in recent years as the spectrum of applications which use this feature has increased. For these systems, it is desirable to create a series of partial bitstreams which represent tasks that can be located in multiple regions in the FPGA substrate. While the transferal of homogeneous collections of lookup-table based logic blocks from region to region has been

shown to be relatively straightforward, it is more difficult to transfer partial bitstreams which contain fixed function resources, such as block RAMs and DSP blocks. To do so, we explore adding enhancements to the FPGA architecture which allow for the migration of partial bitstreams including fixed resources from region to region even if these fixed function resources are not located in the same position in the region. Our approach does not require significant, time-consuming place-and-route during the migration process. We quantify the cost of inserting additional routing resources into the FPGA architecture to allow for easy migration of heterogeneous, fixed function resources. Our experiments show that this flexibility can be added for a relatively low overhead and performance penalty. As mentioned above, the Virtual Bit-Stream (VBS) is a concept of an unfinalized, pre-routed bit-stream which could be loaded almost anywhere on a custom FPGA logic fabric. Unlike classical bit-streams, the VBS is not tied to a specific location on the circuit, hence its "virtual" qualifier. The goal is to generate a single VBS only once for each and every possible location of the logic fabric in the FPGA in a unfinished manner: the time-consuming packing, place and route steps are done offline and only local routing is done at runtime in order to ensure fast decoding time as well as low memory overhead. The VBS concept is pending for a European patent application.

6.1.4. Power Models of Reconfigurable Architectures

Participants: Robin Bonamy, Daniel Chillet, Olivier Sentieys.

Including a reconfigurable area in complex systems-on-chip is considered as an interesting solution to reduce the area of the global system and to support high performance. But the key challenge in the context of embedded systems is currently the power budget and the designer needs some early estimations of the power consumption of its system. Power estimation for reconfigurable systems is a difficult issue since several parameters need to be taken into account to define an accurate model.

One first parameter concerns the choice of tasks to execute and their allocation in the computing resources. Indeed, several hardware implementations of an algorithm can be obtained and exploited by the operating system for a flexible allocation of tasks to optimize energy consumption. These different hardware implementations can be obtained by varying the parallelism level, which has a direct impact on area and execution time and therefore on power and energy consumption. To highlight this point, we made several evaluations of delay, area, power, and energy impacts of loop transformations using High Level Synthesis tools. Real power measurements have been made on an FPGA platform and for different task implementations to build a model of energy consumption versus execution time.

Furthermore, we also considered the opportunity of the dynamic reconfiguration, which makes possible to partially reconfigure a specific part of the circuit while the rest of the system is running. This opportunity has two main effects on power consumption. First, thanks to the area sharing ability, the global size of the device can be reduced and the static (leakage) power consumption can thus be reduced. Secondly, it is possible to delete the configuration of a part of the device which reduces the dynamic power consumption when a task is no longer used.

We analyzed the power consumption during the dynamic reconfiguration on a Virtex 5 board. Three models of the partial and dynamic reconfiguration power consumption with different complexity/accuracy tradeoffs are extracted. These models are used in design space exploration to include impact of reconfiguration on energy consumption of a complete system. We proposed a methodology for power/energy consumption modeling and estimation in the context of heterogeneous (multi)processor(s) and dynamically reconfigurable hardware systems. We developed an algorithm to explore all task mapping possibilities for a complete application (e.g. for H264 video coding) with the aim to extract one of the best solutions with respect to the designer's constraints. This algorithm is a step ahead for defining on-line power management strategies to decide which task instances must be executed to efficiently manage the available power using dynamic partial reconfiguration. All these results are presented in the Robin Bonamy's thesis [17]

6.1.5. Real-time Spatio-Temporal Task Scheduling on 3D Architecture

Participants: Quang-Hai Khuat, Quang-Hoa Le, Emmanuel Casseau, Antoine Courtay, Daniel Chillet.

One of the main advantages offered by a three-dimensional system-on-chip (3D SoC) is the reduction of wire length between different blocks of a system, thus improving circuit performance and alleviating power overheads of on-chip wiring. To fully exploit this advantage, an efficient management referring to allocate temporarily the tasks at different levels of the architecture is greatly important.

In the context of 3D SoC, we have developed several spatio-temporal scheduling algorithms for 3D MultiProcessor Reconfigurable System-on-Chip (3DMPRSoC) architectures composed of a multiprocessor layer and an embedded Field Programmable Gate Array (eFPGA) layer with dynamic reconfiguration. These two layers are interconnected vertically by through-silicon vias (TSVs) ensuring tight coupling between software tasks on processors and associated hardware accelerators on the eFPGA. Our algorithms cope with task dependencies and try to allocate communicating tasks close to each other in order to reduce direct communication cost, thus reducing global communication cost.

In the 3DMPRSoC context, our algorithms favor direct communications including: i) point-to-point communication between hardware accelerators on the eFPGA, ii) communication between software tasks through the Network-on-Chip of the multiprocessor layer, and iii) communication between software task and accelerator through TSV. When a direct communication between two tasks occurs, the data are stored in a shared memory placed onto the multiprocessor layer.

Our work in [68] takes all types of communication into consideration and proposes a scheduling and placement strategy of tasks reducing the global communication cost to 17% compared with our previous algorithm based on Pfair. In this work, the eFPGA layer of the 3DMPRSoC is supposed to contain homogeneous partial reconfiguration regions (PRR) and the size of a hardware accelerator is limited by the size of a PRR. To exceed this limitation, we analyzed the Vertex-List Structure (VLS) method for relocating hardware accelerators of various sizes anywhere onto the eFPGA if resources are available. Then, we proposed VLS-BCF algorithm [49] based on VLS that allows for reducing the overall communication cost significantly – up to 24% – compared to classical methods.

6.1.6. Ultra-Low-Power Reconfigurable Controllers

Participants: Vivek D. Tovinakere, Olivier Sentieys, Steven Derrien.

A key concern in the design of controllers in wireless sensor network (WSN) nodes is the flexibility to execute different control tasks for managing resources, sensing and communications tasks of the node. In this paper, low-power flexible controllers for WSN nodes based on reconfigurable microtasks are presented. A microtask is a digital control unit made up of an FSM and datapath. Scalable architectures for reconfigurable FSMs along with variable precision adders in datapath are proposed for flexible controllers. Power gating as a low power technique is considered for low power operation in reconfigurable microtasks by exploiting coarse grain power gating opportunities in FSMs and adders. Gate-level models are applied to analyze energy savings in logic clusters due to power gating. Power estimation results on typical benchmark microtasks show a $2\times$ to $5\times$ improvement in energy efficiency w.r.t a microcontroller at a cost of $5\times$ when compared with a microtask implemented as an ASIC with higher NRE costs [21].

6.2. Compilation and Synthesis for Reconfigurable Platform

6.2.1. Polyhedral-Based Loop Transformations for High-Level Synthesis

Participants: Steven Derrien, Antoine Morvan, Patrice Quinton, Tomofumi Yuki, Mythri Alle.

After almost two decades of research effort, there now exists a large choice of robust and mature C to hardware tools that are used as production tools by world-class chip vendor companies. Although these tools dramatically slash design time, their ability to generate efficient accelerators is still limited, and they rely on the designer to expose parallelism and to use appropriate data layout in the source program. We believe this can be overcome by tackling the problem directly at the source level, using source-to-source optimizing compilers. More precisely, our aim is to study how polyhedral-based program analysis and transformation can be used to address this problem. In the context of the PhD of Antoine Morvan, we have studied how it was possible to improve the efficiency and applicability of nested loop pipelining (also known as nested software

pipelining) in C to hardware tools. Loop pipelining is a key transformation in high-level synthesis tools as it helps maximizing both computational throughput and hardware utilization.

We have first studied how polyhedral based loop transformations (such as coalescing) could be used to improve the efficiency of pipelining small trip-count inner loops [27] and implemented the transformation in the Gecos source to source toolbox. We also have proposed a technique to widen the applicability of loop pipelining to kernels exposing complex dynamic memory access patterns for which compile time dependency analysis techniques cannot be used efficiently. Our approach borrows from the notion of runtime memory disambiguation used in super scalar processors to add a data dependency hazards detection mechanism to the synthesized circuits. The approach has shown promising results and led to a presentation presented at the 50th ACM/IEEE Design Automation Conference [37]. In addition to our work on nested loop pipelining, we also investigated how to extend existing polyhedral code generation techniques to enable the synthesis of fast and area-efficient control-logic. Our approach was implemented in the Gecos framework and presented at the Field Programmable Technology international conference in late 2013 [63].

6.2.2. *Compiling for Embedded Reconfigurable Multi-Core Architectures*

Participants: Steven Derrien, Olivier Sentieys, Maxime Naullet, Antoine Morvan, Tomofumi Yuki, Ali Hassan El-Moussawi.

Current and future wireless communication and video standards have huge processing power requirements, which cannot be satisfied with current embedded single processor platforms. Most platforms now therefore integrate several processing core within a single chip, leading to what is known as embedded multi-core platforms. This trend will continue, and embedded system design will soon have to implement their systems on platforms comprising tens if not hundred of high performance processing cores. Examples of such architectures are the Xentium processor from by Recore or the Kahrisma processor, a radically new concept of morphable processor from Karlsruhe Institute of Technology (KIT). This evolution will pose significant design challenges, as parallel programming is notoriously difficult, even for domain experts. In the context of the FP7 European Project Alma (Architecture-oriented parallelization for high performance embedded Multi-core systems using scilAb), we are studying how to help designers programming these platforms by allowing them to start from a specification in Matlab and/or Scilab, which are widely used for prototyping image/video and wireless communication applications. Our research work in this field revolves around two topics. The first one aims at exploring how floating-point to fixed-point conversion can be performed jointly with the SIMD instruction selection stage to explore performance/accuracy trade-off in the software final implementation. The second one aims at exploring how program transformation techniques (leveraging the polyhedral model and/or based on the domain specific semantics of scilab built-in functions) can be used to enable an efficient coarse grain parallelization of the target application on such multi-core machines [30].

6.2.3. *Numerical Accuracy Analysis and Optimization*

Participants: Olivier Sentieys, Steven Derrien, Romuald Rocher, Pascal Scalart, Tomofumi Yuki, Aymen Chakhari, Gaël Deest.

Most of analytical methods for numerical accuracy evaluation use perturbation theory to provide the expression of the quantization noise at the output of a system. Existing analytical methods do not consider correlation between noise sources. This assumption is no longer valid when a unique datum is quantized several times. In [35], an analytical model of the correlation between quantization noises is provided. The different quantization modes are supported and the number of eliminated bits is taken into account. The expression of the power of the output quantization noise is provided when the correlation between the noise sources is considered. The proposed approach allows improving significantly the estimation of the output quantization noise power compared to the classical approach, with a slight increase of the computation time.

Trading off accuracy to the system costs is popularly addressed as the word-length optimization (WLO) problem. Owing to its NP-hard nature, this problem is solved using combinatorial heuristics. In [56], a novel approach is taken by relaxing the integer constraints on the optimization variables and obtain an alternate noise-budgeting problem. This approach uses the quantization noise power introduced into the system due to

fixed-point word-lengths as optimization variables instead of using the actual integer valued fixed-point word-lengths. The noise-budgeting problem is proved to be convex in the rounding mode quantization case and can therefore be solved using analytical convex optimization solvers. An algorithm with linear time complexity is provided in order to realize the actual fixed-point word-lengths from the noise budgets obtained by solving the convex noise-budgeting problem.

An analytical approach is studied to determine accuracy of systems including unsmooth operators. An unsmooth operator represents a function which is not derivable in all its definition interval (for example the sign operator). The classical model is no longer valid since these operators introduce errors that do not respect the Widrow assumption (their values are often higher than signal power). So an approach based on the distribution of the signal and the noise was proposed. We focused on recursive structures where an error influences future decision (such as Decision Feedback Equalizer). In that case, numerical analysis method (e.g. Newton Raphson algorithm) can be used. Moreover, an upper bound of the error probability can be analytically determined [43]. We also studied the case of Turbo Coder and Decoder to determine data word-length ensuring sufficient system quality.

One of the limitation of analytical accuracy technique is that they are based on a Signal Flow Graph Representation of the system to be analyzed. This SFG model is currently built-out of a source program by flattening its whole control-flow (including full loop unrolling) which raises significant accuracy analysis issues. In 2013 we have started studying how we could bridge numerical analysis techniques with more compact polyhedral program representations to provide a more general and scalable framework.

6.2.4. Design Tools for Reconfigurable Video Coding

Participants: Emmanuel Casseau, Hervé Yviquel.

In the field of multimedia coding, standardization recommendations are always evolving. To reduce design time taking benefit of available SW and HW designs, Reconfigurable Video Coding (RVC) standard allows defining new codec algorithms. The application is represented by a network of interconnected components (so called actors) defined in a modular library and the behaviour of each actor is described in the specific RVC-CAL language. Dataflow programming, such as RVC applications, express explicit parallelism within an application. However general purpose processors cannot cope with both high performance and low power consumption requirements embedded systems have to face. We have investigated the mapping of RVC applications onto a dedicated multiprocessor platform. Actually, our goal is to propose an automated co-design flow based on the RVC framework. The designer provides the application description in the RVC-CAL language, after which the co-design flow automatically generates a network of processors that can be synthesized on FPGA platforms. The processors are based on a low complexity and configurable TTA processor (Very Long Instruction Word -style processor). The architecture model of the platform is composed of processors with their local memories, an interconnection network and shared memories. Both shared and local memories are used to limit the traditional memory bottleneck. Processors are connected together through the shared memories. The design flow is implemented around two open-source toolsets: Orcc (Open RVC-CAL Compiler: http://orcc.sourceforge.net) and TCE (TTA-based Co-design Environment: http://tce.cs.tut.fi). The inputs of the design flow are the RVC application, the platform configuration (i.e. the configuration of the TTA processors and their number), and the mapping specification (i.e. the mapping of the actors onto the processors). Orcc generates a high-level description of the processors, an intermediate representation of the software code associated to each actor, and the processor interconnection requirements. Then TCE uses these informations to generate a complete multi-processor platform design: the VHDL descriptions of the processors using a pre-existing database of hardware components and the executable binary code that will execute the actors on the processors.

This work is done in collaboration with Mickael Raulet from IETR INSA Rennes and has been implemented in the Orcc open-source compiler and with Jarmo Takala team from Tampere University of Technology (Finland) who is involved in the TCE toolset.

6.3. Interaction between Algorithms and Architectures

6.3.1. Design Methodologies for Software Defined Radios

Participants: Matthieu Gautier, Olivier Sentieys, Emmanuel Casseau, Arnaud Carer, Ganda-Stéphane Ouedraogo, Mai-Thanh Tran, Vaibhav Bhatnagar.

Software Defined Radio (SDR) is a flexible signal processing architecture with reconfiguration capabilities that can adapt itself to various air interfaces. It was first introduced by Joseph Mitola as an underlying structure for Cognitive Radio (CR). The FPGA (Field Programmable Gate Array) technology is expected to play a key role in the development of SDR platforms. FPGA-based SDR is a quite old paradigm and we are fronting this challenge while leveraging the nascent High Level Synthesis tools and languages.
Actually, our goal is to propose methods and tools for rapid implementation of new waveforms in the stringent flexibility paradigm. We proposed a novel design flow for FPGA-based SDR applications [38] [70]. This flow relies upon HLS principles and its entry point is a Domain-Specific Language (DSL) which partly handles the complexity of programming an FPGA and integrates SDR features.

6.3.2. Adaptive Precision under Performance Constraints in OFDM Wireless Receivers

Participants: Olivier Sentieys, Matthieu Gautier, Fernando Cladera [Master's Student].

To cope with rapid variations of channel parameters, wireless receivers are designed with a significant performance margin to reach a given Bit Error Rate (BER), even for worst-case channel conditions. Significant energy savings come from varying at run time processing bit-width, based on estimation of channel conditions, without compromising BER constraints. To validate the energy savings, the energy consumption of basic operators has been obtained from real measurements for different bit-widths on an FPGA and an ARM processor using soft SIMD. Results show that up to 66% of the dynamic energy consumption can be saved using this adaptive technique.

6.3.3. MIMO Systems and Cooperative Strategies for Low-Energy Wireless Networks

Participants: Olivier Berder, Olivier Sentieys, Pascal Scalart, Matthieu Gautier, Le-Quang-Vinh Tran, Duc-Long Nguyen [Master's Student], Ruifeng Zhang, Viet-Hoa Nguyen.

Since a couple of years, the CAIRN team has reached a significant expertise in multi-antenna systems, especially in linear precoding. In order to obtain an efficient, simple and general form of precoders, we considered an SNR-like matrix to approximate the minimum distance. The precoding matrix is first parameterized as the product of a diagonal power allocation matrix and an input-shaping matrix and demonstrated that the minimum diagonal entry of the latter is obtained when the input-shaping matrix is a DFT-matrix. The major advantage of this design is that the solution can be available for all rectangular QAM-modulations and for any number of datastreams [28]. On the other hand the sphere decoder was applied at the receiver side instead of maximum likelihood and the performance complexity trade-off was investigated. Some adjustments of traditional sphere decoding algorithm were mandatory to adapt to the precoded MIMO systems [55].

Another way to exploit the MIMO diversity, especially in WSN where only one antenna can be supported by limited size devices, is to use space-time codes in a distributed manner. In this context, a new protocol, called fully distributed space-time coded (FDSTC) protocol having information exchange between relays, was proposed and compared with the conventional distributed space-time coded (DSTC) protocol using non-regenerative relays (NR-relays) and regenerative relays (R-relays). At the same spectral efficiency, FDSTC has better performance in terms of outage probability in high SNR regions. In terms of energy efficiency, the FDSTC protocol is shown to outperform DSTC for long-range transmissions [32]. As very few dedicated MAC protocols exist, we investigated a novel low-latency MAC protocol (ARQ-CRI) for low-power cooperative wireless sensor networks WSNs, while preserving (in high traffic mode) or even increasing (in low traffic mode) energy-efficiency [54]. An energy efficient opportunistic MAC protocol with the mechanisms of reservation and a relay candidate coordination were also proposed, and the multi-relay transmission probability was analyzed. Simulation and experiment results on a real wireless sensor network platform in different channels demonstrated the proposed scheme greatly reduces the multi-relay transmission probability and achieves about 84% improvement of energy efficiency compared with the traditional opportunistic MAC schemes [66].

6.3.4. Energy Harvesting and Adaptive Wireless Sensor Networks

Participants: Olivier Berder, Olivier Sentieys, Arnaud Carer, Mahtab Alam, Ruifeng Zhang, Trong-Nhan Le.

As tiny sensor nodes are equipped with limited battery, the optimization of the power consumption of these devices is extremely vital. In typical WSN platforms, the radio transceiver consumes major proportion of the energy. Major concerns are therefore to decrease both the transmit power and radio activity. We designed an adaptive transmit power optimization technique that is applied under varying channel to reduce the energy per successful transmitted bit. Each node locally adapts its output power according to the signal-to-noise ratio (SNR) variations (for all the neighbor nodes). It is found that by dynamically adapting the transmit power on average can help to reduce the energy consumption by a factor of two [36].

To further extend the system lifetime of WSN, energy harvesting techniques have been considered as potential solutions for long-term operations. Instead of minimizing the consumed energy as for the case of battery-powered systems, the harvesting node is adapted to Energy Neutral Operation (ENO) to achieve a theoretically infinite lifetime. Several types of energy sources can be used, as light, motion or heat [51]. We even investigated the possibility for a single sediment-microbial fuel cell (MFC) to power a wireless sensor network [31]. Through experiments conducted on the PowWow platform, it was shown that the energy harvesting device adapts to the intermittent power supplied by the MFC, and the radio-transmitter is able to switch from a continuous to degraded mode. Given the harvesting capability, we then tried to design power managers (PM) able to optimize the quality of service of WSN while maintaining ENO. Our PM adapts the duty cycle of the node according to the estimation of harvested energy and the consumed energy provided by a simple energy monitor for a super capacitor based WSN to achieve the ENO [52]. When possible, as is sometimes the case for solar or wind energy, it is also of prime interest to benefit from an accurate energy predictor to estimate the energy that can be harvested in the near future, therefore we proposed a low complexity energy predictor using adaptive filter [53]. Finally, with colleagues from University College of Cork, we recently investigated the possibility to combine energy harvesting platforms with low power wake-up radios. A nano-watt wake-up radio receiver (WUR) was used cooperatively with the main transceiver in order to reduce the wasted energy of idle listening in asynchronous MAC protocols, while still maintaining the same reactivity [50].

6.3.5. Impact of RF Front-End Nonlinearity on WSN Communications.

Participants: Amine Didioui, Olivier Sentieys, Carolynn Bernier [CEA Leti].

In the context of a collaboration with CEA Leti, we studied the impact of RF front-end non-linearity on the performance of wireless sensor networks (WSN). More specifically, we investigated the problem of interference caused by intermodulation between in-band interferers. We analyzed this problem using an enhanced model of signal-to-interference-and-noise ratio (SINR) that includes an interference term due to intermodulation. Using a WSN simulator and the selectivity and the third-order input intercept point (IIP3) specifications of a radio transceiver, we have shown that the new SINR model provides helpful information for the analysis of intermodulation problems caused by in-band signals in IEEE 802.15.4 WSNs. In [45], we presented a reconfigurable receiver model whose purpose is to enable the study of reconfiguration strategies for future energy-aware and adaptive transceivers. This model is based on Figure of Merits of measured circuits. To account for real-life RF interference mechanisms, a link quality estimator is also provided. We show that adapting the receiver performance to the channel conditions can lead to considerable power saving. The models proposed can easily be implemented in a wireless network simulation in order to validate the value of a reconfigurable architecture in real-world deployment scenarios.

6.3.6. HarvWSNet: A Co-Simulation Framework for Energy Harvesting Wireless Sensor Networks.

Participants: Amine Didioui, Olivier Sentieys, Carolynn Bernier [CEA Leti].

Recent advances in energy harvesting (EH) technologies now allow wireless sensor networks (WSNs) to extend their lifetime by scavenging the energy available in their environment. While simulation is the most widely used method to design and evaluate network protocols for WSNs is simulation, existing network simulators are not adapted to the simulation of EH-WSNs and most of them provide only a simple linear battery model. To overcome these issues, we have proposed HarvWSNet, a co-simulation framework based on WSNet and Matlab that provides adequate tools for evaluating EH-WSN lifetime [44]. Indeed, the framework allows for the simulation of multi-node network scenarios while including a detailed description of each node's energy harvesting and management subsystem and its time-varying environmental parameters. A case study based on a temperature monitoring application has demonstrated HarvWSNet's ability to predict network lifetime while minimally penalizing simulation time [40].

6.3.7. Synchronisation Algorithms and Parallel Architecture for Wireless and High-Rate Optical OFDM Systems

Participants: Pramod Udupa, Olivier Sentieys, Arnaud Carer, Pascal Scalart.

Multi-band Coherent Optical OFDM (MB CO-OFDM) is widely predicted to be one of the technologies which will empower 100 Gigabit Ethernet (100GbE) networks. CO-OFDM uses coherent technology and advanced digital signal processing (DSP) to achieve net data rate of 10 Gbps in a single band. This strict throughput requirement puts a constraint on the kind of signal processing algorithms and architectures used for building the system. In [72], a scalable parallel architecture using radix-2^2 for IFFT was proposed. The second proposal consists of a scalable parallel timing synchronization algorithm which can support very high input rates at the receiver. MOPS count as well as area versus throughput for the synchronization algorithm are provided for the OFDM transceiver to show the improvements due to proposed architecture. Architecture exploration was performed using a leading-edge high-level synthesis (HLS) tool.

A novel low complexity parallel algorithm and its associated architecture were proposed for initial synchronization in orthogonal frequency division multiplexing (OFDM) systems. The method is hierarchical and uses auto-correlation for the first step and cross-correlation for the second step [60]. The main advantage of the proposed approach is that it reduces the computational complexity by a factor of five (80%), while achieving similar mean square error (MSE) as cross-correlation based methods. The method uses block-level parallelism for auto-correlation step, which speeds up the computation significantly. After fixed-point analysis, a parallel architecture is proposed to accelerate both coarse and fine synchronization steps. This parallel architecture is scalable and provides speed-up proportional to number of parallel blocks [59].

CELTIQUE Project-Team

5. New Results

5.1. Information Flow Tracking

Participants: Frédéric Besson, Nataliia Bielova, Delphine Demange, Thomas Jensen, David Pichardie.

We investigate different approaches for dynamically tracking information flows.

The first track of work is motivated by web-browser security. In a survey [15], we have classified JavaScript security policies and their enforcement mechanisms in a web-browser. We have identified the problem of stateless web tracking (fingerprinting) and have proposent a novel approach to hybrid information flow monitoring by tracking the knowledge about secret variables using logical formulae. A logic formula quantifies the amount of knowledge stored in a variable. This knowledge representation helps to compare and improve precision of hybrid information flow monitors. We define a generic hybrid monitor parametrised by a static analysis and derive sufficient conditions on the static analysis for soundness and relative precision of hybrid monitors. We instantiate the generic monitor with a combined static constant and dependency analysis. Several other hybrid monitors including those based on well-known hybrid techniques for information flow control are formalised as instances of our generic hybrid mon- itor. These monitors are organised into a hierarchy that establishes their relative precision. The whole framework is accompanied by a formalisation of the theory in the Coq proof assistant [19].

Our second activity is related to SAFE, a clean-slate design for a highly secure computer system, with pervasive mechanisms for tracking and limiting information flows. At the lowest level, the SAFE hardware supports fine-grained programmable tags, with efficient and flexible propagation and combination of tags as instructions are executed. The operating system virtualizes these generic facilities to present an information-flow abstract machine that allows user programs to label sensitive data with rich confidentiality policies. We present a formal, machine-checked model of the key hardware and software mechanisms used to control information flow in SAFE and an end-to-end proof of noninterference for this model in the Coq proof assistant [17].

5.2. Towards efficient abstract domains for regular language based static analysis

Participants: Thomas Genet, Valérie Murat, Yann Salmon.

We develop a specific theory and the related tools for analyzing programs whose semantics is defined using term rewriting systems. The analysis principle is based on regular approximations of infinite sets of terms reachable by rewriting. The tools we develop use, so-called, Tree Automata Completion to compute a tree automaton recognizing a superset of all reachable terms. This over-approximation is then used to prove safety properties on the program by showing that some “bad” terms, encoding dangerous or problematic configurations, are not in the superset and thus not reachable. This is a specific form of, so-called, Regular Tree Model Checking. However, when dealing with infinite-state systems, Regular Tree Model Checking approaches may have some difficulties to represent infinite sets of data. We proposed Lattice Tree Automata, an extended version of tree automata to represent complex data domains and their related operations in an efficient manner. Moreover, we introduce a new completion-based algorithm for computing the possibly infinite set of reachable states in a finite amount of time. This algorithm is independent of the lattice making it possible to seamlessly plug abstract domains into a Regular Tree Model Checking algorithm[27]. As a first instance, we implemented in Timbuk a completion with an interval abstract domain. We shown that this implementation permits to scale up regular tree model-checking of Java programs dealing with integer arithmetics. Now, we aim at applying this technique to the static analysis of programming languages whose semantics is based on terms, like functional programming languages [38].

5.3. Result Certification of Static Program Analysers with Automated Theorem Provers

Participants: Frédéric Besson, Pierre-Emmanuel Cornilleau, Thomas Jensen.

The automation of the deductive approach to program verification crucially depends on the ability to efficiently infer and discharge program invariants. In an ideal world, user-provided invariants would be strengthened by incorporating the result of static analysers as untrusted annotations and discharged by automated theorem provers. However, the results of object-oriented analyses are heavily quantified and cannot be discharged, within reasonable time limits, by state-of-the-art auto- mated theorem provers. In the present work, we investigate an original approach for verifying automatically and efficiently the result of certain classes of object-oriented static analyses using off-the-shelf automated theorem provers. We propose to generate verification conditions that are generic enough to capture, not a single, but a family of analyses which encompasses Java bytecode verification and Fähndrich and Leino type-system for checking null pointers. For those analyses, we show how to generate tractable verification conditions that are still quantified but fall in a decidable logic fragment that is reducible to the Effectively Propositional logic. Our experiments confirm that such verification conditions are efficiently discharged by off-the-shelf automated theorem provers [20].

5.4. Formal Semantics for Multi-threaded Java

Participants: Delphine Demange, Vincent Laporte, David Pichardie.

Recent advances in verification have made it possible to envision trusted implementations of real-world languages. Java with its type-safety and fully specified semantics would appear to be an ideal candidate; yet, the complexity of the translation steps used in production virtual machines have made it a challenging target for verifying compiler technology. One of Java's key innovations, its memory model, poses significant obstacles to such an endeavor. The Java Memory Model is an ambitious attempt at specifying the behavior of multithreaded programs in a portable, hardware agnostic, way. While experts have an intuitive grasp of the properties that the model should enjoy, the specification is complex and not well-suited for integration within a verifying compiler infrastructure. Moreover, the specification is given in an axiomatic style that is distant from the intuitive reordering-based reasonings traditionally used to justify or rule out behaviors, and ill suited to the kind of operational reasoning one would expect to employ in a compiler. We take a step back, and introduces a *Buffered Memory Model* (BMM) for Java [26]. We choose a pragmatic point in the design space sacrificing generality in favor of a model that is fully characterized in terms of the reorderings it allows, amenable to formal reasoning, and which can be efficiently applied to a specific hardware family, namely x86 multiprocessors. Although the BMM restricts the reorderings compilers are allowed to perform, it serves as the key enabling device to achieving a verification pathway from bytecode to machine instructions. Despite its restrictions, we show that it is backwards compatible with the Java Memory Model and that it does not cripple performance on TSO architectures.

5.5. Formal Verification of Static Analysis

Participants: Sandrine Blazy, Martin Bodin, Thomas Jensen, Vincent Laporte, André Oliveira Maroneze, David Pichardie, Alan Schmitt.

Static analyzers based on abstract interpretation are complex pieces of software implementing delicate algorithms. Even if static analysis techniques are well understood, their implementation on real languages is still error-prone.

Using the Coq proof assistant, we formalized of a value analysis (based on abstract interpretation), and a soundness proof of the value analysis. The formalization relies on generic interfaces. The mechanized proof is facilitated by a translation validation of a Bourdoncle fixpoint iterator. The work has been integrated into the CompCert verified C-compiler. Our verified analysis directly operates over an intermediate language of the compiler having the same expressiveness as C. The automatic extraction of our value analysis into OCaml yields a program with competitive results, obtained from experiments on a number of benchmarks and comparisons with the Frama-C tool [21]. The value analysis was applied to a loop bound estimation tool for WCET analysis [22] relying also on program slicing and loop bound calculation.

Moreover, we formalized static analyses for logic programming, relying on results about the relative correctness of semantics in different styles; forward and backward, top-down and bottom-up. The results chosen are paradigmatic of the kind of correctness theorems that semantic analyses rely on and are therefore well-suited to explore the possibilities afforded by the application of interactive theorem provers to this task, as well as the difficulties likely to be encountered in the endeavour [29].

We also study the development of certified information flow analyses based on a formal semantics of JavaScript. We have in particular presented a technique for deriving semantic program analyses from a natural semantics specification of the programming language. The technique is based on the pretty-big-step semantics approach applied to a language with simple objects called O'While. We have specified a series of instrumentations of the semantics that makes explicit the flows of values in a program. This leads to a semantics-based dependency analysis, at the core, e.g., of tainting or information flow analyses in software security [32].

5.6. Certified JavaScript Semantics

Participants: Martin Bodin, Alan Schmitt.

JavaScript is the most widely used web language for client-side applications. Whilst the development of JavaScript was initially just led by implementation, there is now increasing momentum behind the ECMA standardisation process. The time is ripe for a formal, mechanised specification of JavaScript, to clarify ambiguities in the ECMA standards, to serve as a trusted reference for high-level language compilation and JavaScript implementations, and to provide a platform for high-assurance proofs of language properties. We present JScert, a formalisation of the current ECMA standard in the Coq proof assistant, and JSref, a reference interpreter for JavaScript extracted from Coq to OCaml. We give a Coq proof that JSref is correct with respect to JScert and assess JSref using test262, the ECMA conformance test suite. Our methodology ensures that JScert is a comparatively accurate formulation of the English standard, which will only improve as time goes on. We have demonstrated that modern techniques of mechanised specification can handle the complexity of JavaScript [25], [24].

5.7. Concurrent Reversibility

Participant: Alan Schmitt.

Concurrent reversibility has been studied in different areas, such as biological or dependable distributed systems. However, only "rigid" reversibility has been considered, allowing to go back to a past state and restart the exact same computation, possibly leading to divergence. We present a concurrent calculus featuring *flexible reversibility*, allowing the specification of alternatives to a computation to be used upon rollback. Alternatives in processes of this calculus are attached to messages. We show the robustness of this mechanism by encoding more complex idioms for specifying flexible reversibility, and we illustrate the benefits of our approach by encoding a calculus of communicating transactions [30].

5.8. Non linear analysis: fast inference of polynomial invariants

Participants: Thomas Jensen, David Cachera, Arnaud Jobin.

We have proposed an abstract interpretation based method for inferring polynomial invariants.Our analysis uses a form of weakest precondition calculus which was already observed to be well adapted to polynomial disequality guards, and which we extend to equality guards by using parameterized polynomial division. We have shown that the choice of a suitable division operation is crucial at each iteration step in order to compute the invariant. Based on this analysis, we have designed a constraint-based algorithm for inferring polynomial invariants. We have identified heuristics to solve equality constraints between ideals, and implemented the whole analysis algorithm in Maple.A salient feature of this analysis, which distinguishes it from the approaches proposed so far in the literature, is that it does not require the use of Gröbner base computations, which are known to be costly on parameterized polynomials. Our benchmarks show that our analyzer can successfully infer invariants on a sizeable set of examples, while performing two orders of magnitude faster than other existing implementations [16].

CIDRE Project-Team

6. New Results

6.1. Intrusion Detection

6.1.1. Intrusion Detection based on an Analysis of the Flow Control

In 2013, we continue to strengthen our research efforts around intrusion detection parameterized by a security policy.

In [33], we have proposed a language for specifying and composing fine-grained information flow policies. The language used a XML-syntax and has a formal semantic. BSPL enables to precisely specify the expected behavior of applications relatively to their sensitive pieces of information. More precisely it permits to specify where a piece of data owned by an application is allowed to disseminate: in which files or processes.

In [25], we have experimented the previous language (BSPL). We have developed a policy manager for android devices. The manager is able to check the consistency of a policy and to compose two consistent policies. We have also proposed a semi-automatic method for computing information flow policies of applications. We have thus computed some examples of policies and shown that these policies are rich enough to permit benign execution of an application without raising useless alerts and sufficiently restrictive to detect malicious actions induced by a malware.

In [40], we have proposed a new data-structure called System Flow Graph (or SFG in short) that offers a compact representation of how pieces of data flow inside a system. For a given application, the system flow graph describes its external behavior. We have shown that this new data structure suits to represent malware behavior and permits to give an early diagnostic in case of intrusion.

In [36] we have collaborated with Mathieu Jaume from Université de Paris 6 describes a formal framework to draw a correspondence between two types of policy definitions - policies that are defined by properties over states of a system and those that are described by properties over executions of a system.

In [34] and in C.Hauser's PhD desertion, we have extended previous work on kBlare (an IDS that detect illegal flows of information at the kernel level) so as to follow information flows at the network level. To that end, a set of nodes administrated by a single entity can be configured according to a distributed security policy expressed in terms of legal information flows. The different operating systems (kBlare) at each node cooperate by tagging each network packet with a tag that describes the information content of the payload. This way, it is possible to detect illegal information flow of information at the network level. This can be used to detect attacks against confidentiality or integrity of the overall system.

6.1.2. Terminating-Insensitive Non-Interference Verification based on an Information Flow Control

In 2010-2011, we started an informal collaboration with colleagues from CEA LIST laboratory. In 2012, this collaboration has turn into a reality by the funding of a PhD student (Mounir Assaf). This PhD thesis is about the verification of security properties of programs written in an imperative language with pointer aliasing (a subset of C language) by techniques borrowed from the domain of static analysis. One of the property of interest for the security field is called Terminating-Insensitive Non-Interference. Briefly speaking, when verified by a program, this property ensures that the content of any secret variable can not leak into public ones (for any terminating execution). However, this property is too strict in the sense that a large number of programs although perfectly secure are rejected by classical analyzers.

In 2013, Mounir Assaf has studied novel approaches that combine static and dynamic information flow monitoring. These approaches are promising since they enable permissive (accepting a large subset of executions) yet sound (rejecting all insecure executions) enforcement of non-interference. We have investigated a dynamic information flow monitor for a language supporting pointers. Our flow-sensitive monitor relies on prior static analysis in order to soundly enforce non-interference. We have also proposed a program transformation that preserves the behavior of initial programs and soundly inlines our security monitor. This program transformation enables both dynamic and static verification of non-interference in a language supporting pointers. This work has been published in [27] and [45].

6.1.3. Visualization of Security Events

The studies that were performed last year clearly showed that there was an important need for technologies that would allow analysts to handle in a consistent way the various types of log files that they have to study in order to detect intrusion or to perform forensic analysis. Consequently, we proposed this year ELVis, a security-oriented log visualization system that allows the analyst to import its log files and to obtain automatically a relevant representation of their content based on the type of the fields they are made of. First, a summary view is proposed. This summary displays in an adequate manner each field according to its type (i.e. categorical, ordinal, geographical, etc.). Then, the analyst can select one or more fields to obtain some details about it. A relevant representation is then automatically selected by the tool according to the types of the fields that were selected.

ELVis [35] has been presented in VizSec 2013 (part of Vis 2013) in October in Atlanta. A working prototype is currently being tuned in order to perform field trials with our partners in DGA-MI. Next year, we are planing to perform research on how various log files can be combined in the same representation. In the PANOPTESEC project, we will also perform some research on visualization for security monitoring in the context of SCADA systems.

6.2. Privacy

6.2.1. Geoprivacy

With the advent of GPS-equipped devices, a massive amount of location data is being collected, raising the issue of the privacy risks incurred by the individuals whose movements are recorded. In [31], we focus on a specific inference attack called the de-anonymization attack, by which an adversary tries to infer the identity of a particular individual behind a set of mobility traces. More specifically, we propose an implementation of this attack based on a mobility model called Mobility Markov Chain (MMC). A MMC is built out from the mobility traces observed during the training phase and is used to perform the attack during the testing phase. We design several distance metrics quantifying the closeness between two MMCs and combine these distances to build de-anonymizers that can re-identify users in an anonymized geolocated dataset. Experiments conducted on real datasets demonstrate that the attack is both accurate and resilient to sanitization mechanisms such as downsampling. This paper has received the IEEE best student paper award at the conference TrustCom 2013.

In [30], we propose to adopt the MapReduce paradigm in order to be able to perform a privacy analysis on large scale geolocated datasets composed of millions of mobility traces. More precisely, we design and implement a complete MapReduce-based approach to GEPETO. GEPETO (for GEoPrivacy-Enhancing TOolkit) is a flexible software that can be used to visualize, sanitize, perform inference attacks and measure the utility of a particular geolocated dataset. The main objective of GEPETO is to enable a data curator (e.g., a company, a governmental agency or a data protection authority) to design, tune, experiment and evaluate various sanitization algorithms and inference attacks as well as visualizing the following results and evaluating the resulting trade-off between privacy and utility. Most of the algorithms used to conduct an inference attack (such as sampling, k-means and DJ-Cluster) represent good candidates to be abstracted in the MapReduce formalism. These algorithms have been implemented with Hadoop and evaluated on a real dataset. Preliminary results show that the MapReduced versions of the algorithms can efficiently handle millions of mobility traces.

6.2.2. Privacy-enhanced Social Networks

In [38], we have proposed a systematic methodology for evaluating the quality of the privacy proposed by a social networking platform. It is based on an analysis grid organizing a correspondence between a number of design features and properties having an impact on privacy, and a level of distribution. For each property, we consider three possible distribution levels: centralized, decentralized and fully decentralized. For security properties, in particular, we have defined those distribution levels with the help of three different attacker models: an attacker has the ability to compromise either one entity in the system, a pre-defined subset of entites in the system, or the whole set of peers in the system. We argument on the idea that the more powerful the attacker model needed to compromise a property for all users in the system, the higher the privacy level linked to this property. A formal evaluation tool based on lattice structures is then proposed to compare social network systems based on this analysis grid. An example evaluation is also provided, with the thorough analysis of several well-known systems of various kinds, notably leading to the conclusion that some privacy-oriented social networking architectures, presented by their authors as fully distributed, showed centralized characteristics for many privacy-related properties.

6.2.3. Privacy Enhancing Technologies

The development of NFC-enabled smartphones has paved the way to new applications such as mobile payment (m-payment) and mobile ticketing (m-ticketing). However, often the privacy of users of such services is either not taken into account or based on simple pseudonyms, which does not offer strong privacy properties such as the unlinkability of transactions and minimal information leakage. In [26], we introduce a lightweight privacy-preserving contactless transport service that uses the SIM card as a secure element. Our implementation of this service uses a group signature protocol in which costly cryptographic operations are delegated to the mobile phone.

6.2.4. Privacy and Web Services

We have proposed [55] a new model of security policy based for a first part on our previous works in information flow policy and for a second part on a model of Myers and Liskov. This new model of information flow serves web services security and allows a user to precisely define where its own sensitive pieces of data are allowed to flow through the definition of an information flow policy. A novel feature of such policy is that they can be dynamically updated, which is fundamental in the context of web services that allow the dynamic discovery of services. We have also presented an implementation of this model in a web services orchestration in BPEL (Business Process Execution Language).

6.2.5. Privacy-preserving Ad-hoc Routing

6.2.5.1. Proactive Protocol

In [39], we have proposed a *proactive* ad hoc routing protocol that preserves the anonymity of the source and of the destination of the packet flows, and assures the unlinkability of flows between any pair of participants to local observers and to global attackers to a lesser extend. Our solution is based on OLSR and combines Bloom filters and ephemeral identifiers. More specifically, the routing process allows any node to discover the topology of the ad hoc network. Once such a topology is known, a source node can establish beforehand a path to reach any destination node. To conceal the identity of the source and destination nodes, the path may not be the shortest ones nor terminate at the destination node. Then, by including the ephemeral public identifiers of the intermediate nodes into a Bloom filter, the source node is able to specify the nodes that have to rebroadcast packets. Thus, when receiving a packet, a node has simply to check, using its ephemeral private identifier, whether it has to rebroadcast the packet, without knowing the source, the destination, nor the previous and next hop.

6.2.5.2. Reactive Protocol

In [42], we have proposed a classification of privacy preserving properties that ensure privacy in ad hoc network routing. We also proposed a taxonomy of adversary's model to analyse existing privacy preserving ad hoc routing protocols. To improve these protocols and to try address all privacy preserving properties,

we proposed NoName [42], a novel privacy-preserving ad hoc routing protocol. Based on trapdoor, virtual switching and partially disjoint multipath using Bloom filter, NoName ensures anonytmity of the source, of the destination and of intermediate nodes. It also ensures unlinkability between source and message and between destination and message.

In [43], we have proposed another anonymous *proactive* ad hoc routing protocol, called APART, based on Gentry's fully homomorphic cryptography. Even though this technology is currently quite inefficient from a computational perspective, especially for an application in ad-hoc networks, the protocol APART is merely a proof of concept showing that an anonymous proactive protocol is possible thanks to it. The main idea is that each node maintains a routing table that contains only encrypted data. When a source node want to communicate with a destination node, it cooperates with its neighbors to discover the node that is the next hop to the destination node. This is done in such a way that the source node does not know the entry in its routing table that corresponds to the destination, and the next hop does only know that it has to rebroadcast the messages coming from that source.

6.2.6. Right to be forgotten

The right to be forgotten has become an investigation topic in itself within the field of privacy protection. In [46], we present the joint research project funded by the ministry of justice between our team and researchers in law and sociology, in order to examine the current state, in society and in technology, of the notion of a right to be forgotten, to identify the forthcoming computing tools capable of implementing the notion, and to evaluate the relevance of an autonomous legislation to define it and regulate it. In association with this study and in the light of the identified state-of-the-art, we have proposed in [47] a new technique to implement a right to be forgotten in the manner of a degradation of the quality of published data in time, associated with a fully distributed ephemeral publication technology. We show how this technique could fit various use cases in geosocial networks.

6.3. Trust

Digital reputation mechanisms have indeed emerged as a promising approach to cope with the specificities of large scale and dynamic systems. Similarly to real world reputation, a digital reputation mechanism expresses a collective opinion about a target user based on aggregated feedback about his past behavior. The resulting reputation score is usually a mathematical object (*e.g.* a number or a percentage). It is used to help entities in deciding whether an interaction with a target user should be considered. Digital reputation mechanisms are thus a powerful tool to incite users to behave trustworthily. Indeed, a user who behaves correctly improves his reputation score, encouraging more users to interact with him. In contrast, misbehaving users have lower reputation scores, which makes it harder for them to interact with other users. To be useful, a reputation mechanism must itself be accurate against adversarial behaviors. Indeed, a user may attack the mechanism to increase his own reputation score or to reduce the reputation of a competitor. A user may also free-ride the mechanism and estimate the reputation of other users without providing his own feedback. From what has been said, it should be clear that reputation is beneficial in order to reduce the potential risk of communicating with almost or completely unknown entities. Unfortunately, the user privacy may easily be jeopardized by reputation mechanisms which is clearly a strong argument to compromise the use of such a mechanism. Indeed, by collecting and aggregating user feedback, or by simply interacting with someone, reputation systems can be easily manipulated in order to deduce user profiles. Thus preserving user privacy while computing robust reputation is a real and important issue that we address in our work [51], [23].

6.4. Other Topics Related to Security and Distributed Computing

6.4.1. Network Monitoring and Fault Detection

Monitoring a system consists in collecting and analyzing relevant information provided by the monitored devices, so as to be continuously aware of the system state (situational awareness). However, the ever growing complexity and scale of systems makes both real time monitoring and fault detection a quite tedious task. Thus

the usually adopted option is to focus solely on a subset of information states, so as to provide coarse-grained indicators. As a consequence, detecting isolated failures or anomalies is a quite challenging issue. We propose in [24], [44] to address this issue by pushing the monitoring task at the edge of the network. We present a peer-to-peer based architecture, which enables nodes to adaptively and efficiently self-organize according to their "health" indicators. By exploiting both temporal and spatial correlations that exist between a device and its vicinity, our approach guarantees that only isolated anomalies (an anomaly is isolated if it impacts solely a monitored device) are reported on the fly to the network operator. We show that the end-to-end detection process, *i.e.*, from the local detection to the management operator reporting, requires a logarithmic number of messages in the size of the network.

6.4.2. Metrics Estimation on Very Large Data Streams

In [12], we consider the setting of large scale distributed systems, in which each node needs to quickly process a huge amount of data received in the form of a stream that may have been tampered with by an adversary (*i.e.*, data items ordering can be manipulated by an omniscient adversary [13]). In this situation, a fundamental problem is how to detect and quantify the amount of work performed by the adversary. To address this issue, we propose AnKLe (for Attack-tolerant eNhanced Kullback- Leibler divergence Estimator), a novel algorithm for estimating the KL divergence of an observed stream compared to the expected one. AnKLe combines sampling techniques and information-theoretic methods. It is very efficient, both in terms of space and time complexities, and requires only a single pass over the data stream. Experimental results show that the estimation provided by AnKLe remains accurate even for different adversarial settings for which the quality of other methods dramatically decreases. Considering n as the number of distinct data items in a stream, we show that AnKLe is an (ε, δ)-approximation algorithm with a space complexity $\widetilde{\mathcal{O}}(\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2})$ bits in "most" cases, and $\widetilde{\mathcal{O}}(\frac{1}{\varepsilon} + \frac{n-\varepsilon^{-1}}{\varepsilon^2})$ otherwise. To the best of our knowledge, an approximation algorithm for estimating the Kullback-Leibler divergence has never been analyzed before. We go a step further by considering in [21] the problem of estimating the distance between any two large data streams in small-space constraint. This problem is of utmost importance in data intensive monitoring applications where input streams are generated rapidly. These streams need to be processed on the fly and accurately to quickly determine any deviance from nominal behavior. We present a new metric, the *Sketch ☆-metric*, which allows to define a distance between updatable summaries (or sketches) of large data streams. An important feature of the *Sketch ☆-metric* is that, given a measure on the entire initial data streams, the *Sketch ☆-metric* preserves the axioms of the latter measure on the sketch (such as the non-negativity, the identity, the symmetry, the triangle inequality but also specific properties of the f-divergence or the Bregman one). Extensive experiments conducted on both synthetic traces and real data sets allow us to validate the robustness and accuracy of the *Sketch ☆-metric*.

6.4.3. Robustness Analysis of Large Scale Distributed Systems

In the continuation of [53] which proposed an in-depth study of the dynamicity and robustness properties of large-scale distributed systems, in [22], we analyze the behavior of a stochastic system composed of several identically distributed, but non independent, discrete-time absorbing Markov chains competing at each instant for a transition. The competition consists in determining at each instant, using a given probability distribution, the only Markov chain allowed to make a transition. We analyze the first time at which one of the Markov chains reaches its absorbing state. When the number of Markov chains goes to infinity, we analyze the asymptotic behavior of the system for an arbitrary probability mass function governing the competition. We give conditions for the existence of the asymptotic distribution and we show how these results apply to cluster-based distributed systems when the competition between the Markov chains is handled by using a geometric distribution.

6.4.4. Secure Uniform Sampling in Dynamic Systems

In [21], we consider the problem of achieving uniform node sampling in large scale systems in presence of a strong adversary. We first propose an omniscient strategy that processes on the fly an unbounded and arbitrarily biased input stream made of node identifiers exchanged within the system, and outputs a stream that preserves Uniformity and Freshness properties. We show through Markov chains analysis that both properties hold

despite any arbitrary bias introduced by the adversary. We then propose a knowledge-free strategy and show through extensive simulations that this strategy accurately approximates the omniscient one. We also evaluate its resilience against a strong adversary by studying two representative attacks (flooding and targeted attacks). We quantify the minimum number of identifiers that the adversary must insert in the input stream to prevent uniformity. To our knowledge, such an analysis has never been proposed before.

DIONYSOS Project-Team

6. New Results

6.1. Quality of Experience

Participants: Yassine Hadjadj-Aoul, Adlen Ksentini, Gerardo Rubino, César Viho, Pantelis Frangoudis, Hyunhee Park, Kandaraj Piamrat.

We continue the development of the PSQA technology (Pseudo-Subjective Quality Assessment) in the area of Quality of Experience (QoE). PSQA is today a stable technology allowing to build measuring modules capable of quantifying the quality of a video or an audio sequence, as perceived by the user, when received through an IP network. It provides an accurate and efficiently computed evaluation of quality. Accuracy means that PSQA gives values close to those that can be obtained from a panel of human observers, under a controlled subjective testing experiment, following an appropriate standard (which depends on the type of sequence or application). Efficiency means that our measuring tool can work in real time, if necessary. Observe that perceived quality is, in general, the main component of QoE when the application or service involves video and audio, or voice. PSQA works by analyzing the networking environment of the communication and some the technical characteristics of the latter. It works without any need to the original sequence (as such, it belongs to the family of *no-reference* techniques).

It must be pointed out that a PSQA measuring or monitoring module is network-dependent and application-dependent. Basically, for each specific networking technology, application, service, the module must be built from scratch. But once built, it works automatically and efficiently, allowing if necessary its use in real time, typically for controlling purposes.

Learning tools. At the heart of the PSQA approach there is the statistical learning process necessary to develop measuring modules. So far we have been using Random Neural Networks (RNNs) for that purpose (see [74] for a general description), but recently, we started to explore other approaches. For instance, in the last ten years a new computational paradigm was presented under the name of *Reservoir Computing* (RC) [71] with the goal of attacking the main limitations in training time for recurrent neural networks while introducing no significant disadvantages. Two RC models have been proposed independently and simultaneously under the name of *Liquid State Machine* (LSM) [73] and *Echo State Networks* (ESN) [71]. They constitute today one of the basic paradigms for Recurrent Neural Networks modeling [72]. The main characteristic of the RC model is that it separates two parts: a static sub-structure called *reservoir* which involves the use of cycles in order to provide dynamic memory in the network, and a parametric part composed of a function such as a multiple linear regression or a classical single layer network. The reservoir can be seen as a high-dimensional dynamical system that expand the input stream in a space of states. The learning part of the model is the parametric one. In [41] we propose a new learning tool which merges the capabilities of Random Neural Networks (RNNs) with those of RC models. We keep some of the nice features of RNNs with the ability of RC models in predicting time series values. Our tool is called Echo State Queueing Network. In the paper, we illustrate its performances in predicting, in particular, Internet traffic. In [63], more results about the good behavior of our new tool are presented.

QoE for SVC. A recent video encoding scheme called Scalable Video Coding (SVC) provides the flexibility and the capability to adapt the video quality to varying network conditions and heterogeneous users. Last year, we started to look at the relations between the way SVC is used and the obtained perceived quality. This year we continued these efforts, together with exploring the use of QoE estimation tools for SVC video coding in network control. In [46] we evaluate different configurations for SVC-based adaptive streaming in terms of user QoE. The aim is to provide recommendations about the different rates to be used in order to create the video representation configuration. These results are part of the PhD [11]. In [25], we extended our previous work on SVC in DVB-T2, by proposing an analytical model to evaluate the performance of associating SVC with DVB-T2 and QoE. To do this, we developed a discrete time Markov Chain model which captures the

system evolution in terms of number of SVC layers that need to be decoded in order to increase user QoE. In [45], we introduced a new solution to be used by a DASH client for selecting the video representation. Our proposal relies on using the PTP synchronization protocol in order to estimate the end-to-end delays between the client and the server, and hence to correlate this information with network load. The correlation between delays and load was based on a fitting function.

In [54], we focus on SVC multicast over IEEE 802.11 networks. Traditionally, multicast uses the lowest modulation resulting in a video with only base quality even for users with good channel conditions. To optimize QoE, we propose to use multiple multicast sessions with different transmission rates for different SVC layers. The goal is to provide at least the multicast session with acceptable quality to users with bad channel conditions and to provide additional multicast sessions having SVC enhancement layers to users with better channel conditions. The selection of modulation rate for each SVC layer and for each multicast session is achieved with binary integer linear programming depending on network conditions with a goal to maximize global QoE. The results show that our algorithm maximizes global QoE by providing highest quality videos to users with good channel conditions and by guaranteeing at least acceptable QoE for all users.

VoIP. We continued to work on the perceptual quality of voice-based applications and services. In [17], we consider a well-known and widely used *full-reference* technique for measuring speech quality called PESQ, and we propose a learning-based tool for approximating PESQ output without any need for the original signal, following the same black-box parametric PSQA approach. The procedure uses the Echo State Networks previously mentioned.

In [48], we propose a new packet loss model that differentiates loss instances depending on their perceptual impact. In particular, the model captures the differences between short and long interruptions from the perceptual quality viewpoint. In some cases, the delays and their variation have a strong impact on the perceived quality. In [49] we explore the variability of packet delays on MANETs. For that purpose, a wide range of representative scenarios are defined and simulated. The gathered traces are then inspected from qualitative and quantitative perspectives. In [50], a Markovian model is proposed to capture these and other features of delays in the same class of mobile networks.

6.2. Network Economics

Participant: Bruno Tuffin.

The general field of network economics, analyzing the relationships between all actors of the digital economy, has been an important subject for years in the team.

A new book on the subject. We have published a book on this broad topic [61]. Presenting a balance of theory and practice, this up-to-date guide provides a comprehensive overview of the key issues in telecommunication network economics, as well as the mathematical models behind the solutions. These mathematical foundations enable the reader to understand the economic issues arising at this pivotal time in network economics, from business, research, and political perspectives. This is followed by a unique practical guide to current topics, including app stores, volume-based pricing, auctions for advertisements, search engine business models, the network neutrality debate, the relationship between mobile network operators and mobile virtual network operators, and the economics of security. The guide discusses all types of players in telecommunications, from users, to access and transit network providers; to service providers (including search engines, cloud providers or content delivery networks); to content providers, and regulatory bodies. The book is designed for graduate students, researchers, and industry practitioners working in telecommunications.

Research contributions in network economics during 2013 can be decomposed into the application of auction theory, cognitive networks, and network/search neutrality analysis.

Auction theory. In the next generation Internet, we have seen the convergence of multimedia services and Internet with the mobility of users. Vertical handover decision (VHD) algorithms are essential components of the mobility management architecture in mobile wireless networks. VHD algorithms help mobile users to choose the best mobile network to connect among available candidates. It also can help the network manager to optimize easily the limited resources shared among the network providers and the users. In [26], we formulate

VHD algorithm as a resource allocation problem for down-link transmission power in multiple W-CDMA networks and show how combinatorial double-sided auctions can be applied to this specific problem. The proposed pricing schemes make use of the signal interference to noise ratio (SINR), achievable data rates, power allocation at mobile networks, and monetary cost as decision criteria, and our model differentiates new calls and on-going communications to take into account that the last category has somewhat more importance. Several combinatorial double-sided auction are proposed to maximize the social welfare and /or to provide incentives for mobile users and mobile operators to be truth-telling in terms of valuation or cost. Finally, the economic properties of the different proposed pricing schemes are also studied by means of simulations.

Cognitive networks. Cognitive radio technologies for spectrum sharing have received an enormous interest from the research community for the last decade, and more recently from regulators and mobile operators. We have studied a cognitive radio network in [47] where primary operator and an entrant secondary operator compete for users. The system is modeled using queueing and game theories. The economic viability of supporting the secondary operator service using an opportunistic access to the spectrum owned by the primary operator is assessed. Against the benchmark of the primary operator operating as a monopolist, we show that the entry of the secondary operator is desirable from an efficiency perspective, since the carried traffic increases. For a range of parameter values, a lump sum payment can be designed so that the incumbent operator has an incentive to let the secondary operator enter. Additionally, the opportunistic access setting has been compared against a leasing-based alternative, and we have concluded that the former outperforms the latter in terms of efficiency and incentive.

Network/search neutrality analysis. Network neutrality is the topic of a vivid and very sensitive debate, in both the telecommunication and political worlds, because of its potential impact in everyday life. That debate has been raised by Internet Service Providers (ISPs), complaining that content providers (CPs) congest the network with insufficient monetary compensation, and threatening to impose side payments to CPs in order to support their infrastructure costs. While there have been many studies discussing the advantages and drawbacks of neutrality, there is no game-theoretical work dealing with the observable situation of competitive ISPs in front of a (quasi-)monopolistic CP. However, this is a typical situation that is condemned by ISPs, and, according to them, another reason of the non-neutrality need. We develop and analyze in [23] a model describing the relations between two competitive ISPs and a single CP, played as a three-level game corresponding to three different time scales. At the largest time scale, side payments (if any) are determined. At a smaller time scale, ISPs decide their (flat-rate) subscription fee (toward users), then the CP chooses the (flat-rate) price to charge users. Users finally select their ISP (if any) using a price-based discrete choice model, and decide whether to also subscribe to the CP service. The game is analyzed by backward induction. As a conclusion, we obtain among other things that non-neutrality may be beneficial to the CP, and not necessarily to ISPs, unless the side payments are decided by ISPs.

The very related recently raised search neutrality debate questions the ranking methods implemented by search engines: when a search is performed, do they (or should they) display the web pages ordered according to the quality-of-experience (relevance) of the content? In [68], we analyze that question in a setting when content is offered for free, content providers making revenue through advertising. For content providers, determining the amount of advertising to add to their content is a crucial strategic decision. Modeling the trade-off between the revenue per visit and the attractiveness, we investigate the interactions among competing content providers as a non-cooperative game, and consider the equilibrium situations to compare the different ranking policies. Our results indicate that when the search engine is not involved with any high-quality content provider, then it is in its best interest to implement a neutral ranking, which also maximizes user perceived quality-of-experience and favors innovation. On the other hand, if the search engine controls some high-quality content, then favoring it in its ranking and adding more advertisement yields a larger revenue. This is not necessarily at the expense of user perceived quality, but drastically reduces the advertising revenues of the other content providers, hence reducing their chances to innovate.

6.3. Wireless and Mobile Networks

Participants: Yassine Hadjadj-Aoul, Adlen Ksentini, César Viho, Osama Arouk, Btissam Er-Rahmadi, Hyunhee Park, Kandaraj Piamrat.

We continue our activities around wireless and mobile networks, where we focus particularly on 4G networks as well as on a new mobile architecture known as mobile cloud.

LTE improvements. First part of our works concentrates on emerging applications and their impact on 4G networks. In [58], we proposed a solution to handle social network traffic, which is characterized by its elasticity and intensity in a short period of time. The proposed contribution is based on content detection systems such as Deep Packet Inspection (DPI) to identify traffic belonging to a group of users (sharing the same content) of a social network. Upon detecting the type of traffic, we proposed to control it by creating a multicast group. This would reduce the amount of traffic exchanged by switching from unicast communications to multicast communications. Another solution is to cache, at the geographically nearest base station, the shared content among users. Here we positioned ourselves in the case where the social network traffic comes from the same geographical region. We also investigated network decentralization in conjunction with the selective IP traffic offload approaches to handle such increased data traffic. We first devised different approaches based on a per-destination-domain-name basis, which offer operators a fine-grained control to determine whether a new IP connection should be offloaded or accommodated via the core network. Two of our solutions are based on Network Address Translation (NAT) named simple-NATing and twice-NATing, whereas a third one employs simple tunneling, and a fourth adopts multiple Access Point Names. We also proposed methods enabling user equipment devices to always have efficient packet data network connections [30]. Another aspect, we addressed is the gateway selection process, where in [59] we argue the need for other metrics to improve the gateway selection mechanisms in distributed mobile networks. We therefore proposed to consider the end-to-end connection and the service/application type as two important additional metrics in the selection of data anchor gateways in the context of the Evolved Packet System (EPS).

M2M. In [56], [32] we addressed another type of traffic that appeared these last years, namely Machine to machine communication or Machine Type Communication (MTC). Such traffic is known by its intensity and its impact on increasing congestion in both parts of the 4G networks, the Radio Access Network (RAN) and the core network parts. The main spirit of the proposed solutions is to proactively anticipate system overload by reducing the amount of MTC signaling messages exchanged in normal network operations. The first solution reduces the number of exchanged signaling messages when triggering MTC devices with low mobility. It enables direct triggering of MTC devices with low mobility by MTC-IWF (MTC InterWorking Function), without involving the MME (Mobility Management Entity). Second solution defines a method for controlling and anticipating network overload in case of an event/scenario whereby a mass of messages with some common Information Elements (IE) are to be exchanged on an interface between two nodes. The network overload control is achieved via dynamic creation of a profile characterizing the event/scenario and the common IEs.

Home networks. In-home wireless networks are now wide-spreading as today's home network is composed of at least one wireless network. The dramatic increase of traffic in such networks yields to difficulties in guaranteeing user experience especially for some specific services like IPTV. This is particularly complicated when using UDP at transport layer and traditional MAC protocol at link layer. Therefore, we investigated comparison of different combinations of transport and link layer performances for the delivery of IPTV. For validation, we use NS-3 and a realistic propagation model generated with a real house description. We analyze impact of link layer (with or without coordination) and transport layer (UDP or TCP). Then, we propose a combined solution using TCP over a coordinated MAC protocol (see [52]). The proposed solution can be easily deployed in real products and is compatible with existing devices.

Another part of our activities in wireless network are related to energy saving. Indeed, one of the biggest problem today in the wireless world is that wireless devices are battery driven, which reduce their operating lifetime. The experimental measurements we have achieved in [18], [42] revealed that operating system overhead causes a drop in performance and energy consumption properties as compared to the GPP in case of certain low video qualities. We propose, thus, a new approach for energy-aware processor switching (GPP or DSP) which takes into consideration the video quality. We show the pertinence of our solution in the context of adaptive video decoding and implement it on an embedded Linux operating system.

6.4. Future Networks

Participants: Yassine Hadjadj-Aoul, Adlen Ksentini, Leila Ghazzai, Jean-Michel Sanner.

Mobile cloud. One of the 5G-architecture visions considers the usage of cloud to build mobile networks and help in decentralizing mobile networks on demand, elastically, and in the most cost-efficient way. This concept of carrier cloud becomes of vital importance knowing that several cloud providers are distributing their cloud/network, globally deploying more regional data centers, to meet their ever-increasing business demands. As an important enabler of the carrier cloud concept, network function virtualization (NFV) is gaining great momentum among industries. NFV aims for decoupling the software part from the hardware part of a carrier network node, traditionally referring to a dedicated hardware, single service and single-tenant box, and that is using virtual hardware abstraction. Network functions become thus a mere code, runnable on a particular, preferably any, operating system and on top of a dedicated hardware platform. The ultimate objective is to run network functions as software in standard virtual machines (VMs) on top of a virtualization platform in a general-purpose multi-service multi-tenant node (e.g., Carrier Grade Blade Server) put into the cloud. In [31], we presented and detailed the Follow Me Cloud (FMC) concept, whereby mobile services hosted in federated clouds follow mobile users as they move and according to their needs. We then provided in [55] a detailed analytical model based on continuous time Markov chain which considers to evaluate the performance of FMC in terms of service migration cost and QoS gain for user. An efficient mobile cloud cannot be built without efficient algorithms for the placement of NFV over this federated cloud. In this vein, in [57] we argued the need for avoiding or minimizing the frequency of mobility gateway (S-GW) relocations and discussed how this gateway relocation avoidance can be reflected in an efficient network function placement algorithm for the realization of mobile cloud. The problem was modeled by an Integer Linear Problem and proved to be NP hard. Therefore, two heuristics were proposed for the creation of a NFV S-GW instance in the cloud.

SDN. We started an activity on Software Defined Networking (SDN), a recent idea proposed to handle network management problems. SDN are becoming an important issue with the ever-increasing network complexity. They are proposed as an alternative to the current architecture of the Internet, which cannot meet the supported services requirements such as Quality of Service/Experience (Qos/QoE), security and energy consumption. We particularly address the scalability issue by proposing a hierarchical controller-based architecture handling the whole control chain.

6.5. Interoperability assessment and improvement

Participants: César Viho, Anthony Baire, Nanxing Chen.

The Internet of Things (IoT) brings new challenges to interoperability assessment by introducing the necessity to deal with non reliable environments connecting plenty billions of objects widely distributed. Therefore, in the recent period, we propose an interoperability testing methodology using a *passive* approach. It appeared more suitable for this distributed, unrliable and constrained environment brought by IoT. We have also developed a tool that implements this passive method. It has been used successfully to test CoAP implementations during the two CoAP Plugtest interoperability sessions on IoT protocols (CoAP and 6LoWPAN) organized by ETSI and IPSO Alliance. These contributions are published in [10].

6.6. Performance Evaluation of Distributed Systems

Participants: Bruno Sericola, Romaric Ludinard.

Network Monitoring and Fault Detection. Monitoring a system is the ability of collecting and analyzing relevant information provided by the monitored devices so as to be continuously aware of the system state. However, the ever growing complexity and scale of systems makes both real time monitoring and fault detection a quite tedious task. Thus the usually adopted option is to focus solely on a subset of information states, so as to provide coarse-grained indicators. As a consequence, detecting isolated failures or anomalies is a quite challenging issue. We propose in [38] and [60] to address this issue by pushing the monitoring task at the edge of the network. We present a peer-to-peer based architecture, which enables nodes to adaptively and efficiently self-organize according to their ”health” indicators. By exploiting both temporal and spatial

correlations that exist between a device and its vicinity, our approach guarantees that only isolated anomalies (an anomaly is isolated if it impacts solely a monitored device) are reported on the fly to the network operator. We show that the end-to-end detection process, *i.e.*, from the local detection to the management operator reporting, requires a logarithmic number of messages in the size of the network. These results also led to the patent [70].

Robustness Analysis of Large Scale Distributed Systems. In the continuation of previous work which proposed an in-depth study of the dynamicity and robustness properties of large-scale distributed systems, in [15] we analyze the behavior of a stochastic system composed of several identically distributed, but non independent, discrete-time absorbing Markov chains competing at each instant for a transition. The competition consists in determining at each instant, using a given probability distribution, the only Markov chain allowed to make a transition. We analyze the first time at which one of the Markov chains reaches its absorbing state. When the number of Markov chains goes to infinity, we analyze the asymptotic behavior of the system for an arbitrary probability mass function governing the competition. We give conditions for the existence of the asymptotic distribution and we show how these results apply to cluster-based distributed systems when the competition between the Markov chains is handled by using a geometric distribution.

Secure Uniform Sampling in Dynamic Systems. In [37], we consider the problem of achieving uniform node sampling in large scale systems in presence of a strong adversary. We first propose an omniscient strategy that processes on the fly an unbounded and arbitrarily biased input stream made of node identifiers exchanged within the system, and outputs a stream that preserves Uniformity and Freshness properties. We show through Markov chains analysis that both properties hold despite any arbitrary bias introduced by the adversary. We then propose a knowledge-free strategy and show through extensive simulations that this strategy accurately approximates the omniscient one. We also evaluate its resilience against a strong adversary by studying two representative attacks (flooding and targeted attacks). We quantify the minimum number of identifiers that the adversary must insert in the input stream to prevent uniformity. To our knowledge, such an analysis has never been proposed before.

6.7. Monte Carlo

Participants: Gerardo Rubino, Bruno Tuffin, Pablo Sartor Del Giudice.

We maintain a research activity in different areas related to dependability, performability and vulnerability analysis of communication systems, using both the Monte Carlo and the Quasi-Monte Carlo approaches to evaluate the relevant metrics. Monte Carlo (and Quasi-Monte Carlo) methods often represent the only tool able to solve complex problems of these types. However, when the events of interest are rare, simulation requires a special attention, for two reasins: the need in accelerating the occurrence of those events and in getting unbiased estimators of them with a sufficiently small relative variance. This is the main problem in the area. Dionysos' work focuses then in dealing with the rare event situation. Rare event simulation has been reviewed in [22].

Multidimensional integrals. In [20], we present a versatile Monte Carlo method for estimating multidimensional integrals, with applications to rare-event probability estimation. The method combines two distinct and popular Monte Carlo simulation techniques, Markov chain Monte Carlo and Importance Sampling, into a single algorithm. We show that for some applied numerical examples the proposed Markov Chain Importance Sampling algorithm performs better than methods based solely on Importance Sampling or MCMC.

Static models. Static reliability analysis has been the topic of an extensive activity in the group for years. Exact evaluation of static network reliability parameters belongs to the NP-hard family and Monte Carlo simulation is therefore a relevant tool to provide estimations for them.

In [67], we first review a Recursive Variance Reduction (RVR) estimator which approaches the unreliability metric by recursively reducing the graph from the random choice of the first working link on selected cuts. We show that the method does not verify the bounded relative error (BRE) property as reliability of individual links goes to one, i.e., that the estimator is not robust in general to high reliability of links. We then propose to use the decomposition ideas of the RVR estimator in conjunction with the Importance Sampling technique.

Two new estimators are presented: the first one, called Balanced Recursive Decomposition estimator, chooses the first working link on cuts uniformly, while the second, called Zero-Variance Approximation Recursive Decomposition estimator, tries to mimic the estimator with variance zero for this technique. We show that in both cases the BRE property is verified and, moreover, that a Vanishing Relative Error property can be obtained for the Zero-Variance Approximation RVR under specific sufficient conditions. A numerical illustration of the power of the methods is provided on several benchmark networks.

The same problem is also analyzed in [19] by a novel method that exploits a generalized splitting (GS) algorithm. We show that the proposed GS algorithm can accurately estimate extremely small unreliabilities and we exhibit large examples where it performs much better than existing approaches. Remarkably, it is also flexible enough to dispense with the frequently made assumption of independent edge failures.

On the same type of model, we propose in [51] an adaptive parameterized method to approximate the zero-variance change of measure for the evaluation of static network reliability models, with links subject to failures. The method uses two rough approximations of the unreliability function, conditional on the states of any subset of links being fixed. One of these approximation, based on mincuts, under-estimates the true unknown unreliability, whereas the other one, based on minpaths, over-estimates it. Our proposed change of measure takes a convex linear combination of the two, estimates the optimal (graph-dependent) coefficient in this combination from pilot runs, and uses the resulting conditional unreliability approximation at each step of a dynamic Importance Sampling algorithm. This new scheme is more general and more flexible than a previously-proposed zero-variance approximation one, which is based on mincuts only and which was shown to be robust asymptotically when unreliabilities of individual links decrease toward zero. Our numerical examples show that the new scheme is often more efficient when the unreliabilities of the individual links are not so small but the overall unreliability is small because the system can fail in many ways. Part of these results are in the PhD [13].

In [43], we present a generalization of the above static models to cases for which the component failures are not independent. To model the dependence and also to develop effective simulation methods that estimate the system unreliability, we extend the static model into an auxiliary dynamic one where the components fail at random times, according to a Marshall-Olkin multivariate exponential distribution. We examine and compare different versions of this model and develop efficient unreliability estimation methods based on conditional Monte Carlo and on a generalized splitting methodology.

In [28], a different splitting algorithm is proposed for solving the same static problem, which is converted into a dynamic one by means of the Creation Process of Elperin, Gerbtsbakh and Lomonosov. The classic splitting technique is then applied, and the obtained results are explored through several numerical experiments. The relative error and the covering properties of the obtained estimator are particularly studied.

In [29], a generalization of the basic model is studied using Monte Carlo. The idea is that the system (the network) works when the terminal nodes are connected *by at least one path whose length is less than or equal to a given parameter* d. This is called Diameter Constrained Reliability. If the parameter d is greater than or equal to the longest path in the network (or between terminals), the problem is the classic one. The paper proposes a variance reduction technique for the estimation of the system's reliability in this setting. In [21], we analyze the particular case of $d = 2$ using exact techniques. These results are part of the thesis [14].

Finally, in [34] and [36] we made general presentations on the rare event problem in general, and on some of the team's results concerning the design of efficient techniques to analyze them.

6.8. Analytic models

Participants: Raymond Marie, Bruno Sericola, Gerardo Rubino, Laura Aspirot.

New books about Markovian models and applications. The book [65] is the french version of the book [66]. Markov chains are a fundamental class of stochastic processes. They are the main modeling tool used in our team. They are widely used to solve problems in a large number of domains such as operations research, computer science, communication networks and manufacturing systems. The success of Markov chains is mainly due to their simplicity of use, the large number of available theoretical results and the quality

of algorithms developed for the numerical evaluation of many metrics of interest. The books present the theory of both discrete-time and continuous-time homogeneous Markov chains. They examine the explosion phenomenon, the Kolmogorov equations, the convergence to equilibrium and the passage time distributions to a state and to a subset of states. These results are applied to birth-and-death processes. A detailed study of the uniformization technique by means of Banach algebra results is also developed. This technique is used for the transient analysis of several queuing systems.

Another book entitled "Markov Chains and Dependability Theory" will be published soon by Cambridge University Press (see http://www.amazon.fr/Markov-Chains-Dependability-Theory-Gerardo/dp/1107007577/). Dependability metrics are omnipresent in every engineering field, from simple ones through to more complex measures combining performance and dependability aspects of systems. The book presents the mathematical basis of the analysis of these metrics in the most used framework, Markov models, describing both basic results and specialised techniques. It presents both discrete and continuous time Markov chains before focusing on dependability measures, which necessitate the study of Markov chains on subsets of states representing different user satisfaction levels for the modelled system. Topics covered include Markovian state lumping, analysis of sojourns on subset of states of Markov chains, analysis of most dependability metrics, fundamentals of performability analysis, and bounding and simulation techniques designed to evaluate dependability measures. The book is of interest to graduate students and researchers in all areas of engineering where the concepts of lifetime, repair duration, availability, reliability and risk are important.

Fluid models. In [53] and [44] we propose a new way of transporting video flows on a peer-to-peer architecture of the Bit-Torrent type. We analyze the performance obtained by our proposal by means of fluid views of the systems, that is, by representing them using differential equations. In [53] the basic idea is to select the downloading peers according to their progress in the downloading process: a given peer only sends chunks to other peers that are downloading at least roughly in the same "area" of the stream. The system is improved in [44] where the main resource (the available bandwidth) is distributed differently among the peers, giving some kind of priority to those nodes remaining more time connected.

In [39], we look at the problem of approximating Markovian views of the Machine Repaiman Model where life-times and repair times have Phase-type distributions, by differential equations. The machine population goes to infinity, and we analyze the properties of the limiting differential equation (once the Markovian sequence of models is properly scaled) and their relations with the initial models. In [63] we describe these results and other results concerning the same type of limiting processes, but concerning peer-to-peer networks. We discuss here the convergence aspects; the properties of the fluid models themselves are discussed in the two papers [53] and [44] mentioned before.

DREAM Project-Team

6. New Results

6.1. Diagnosis of large scale discrete event systems

Participants: Marie-Odile Cordier, Christine Largouët, Sophie Robin, Laurence Rozé, Yulong Zhao.

The problem we deal with is monitoring complex and large discrete-event systems (DES) such as an orchestration of web services or a fleet of mobile phones. Two approaches have been studied in our research group. The first one consists in representing the system model as a discrete-event system by an automaton. In this case, the diagnostic task consists in determining the trajectories (a sequence of states and events) compatible with the sequence of observations. From these trajectories, it is then easy to determine (identify and localize) the possible faults. In the second approach, the model consists in a set of predefined characteristic patterns. We use temporal patterns, called chronicles, represented by a set of temporally constrained events. The diagnostic task consists in recognizing these patterns by analyzing the flow of observed events.

6.1.1. Distributed monitoring with chronicles - Interleaving diagnosis and repair - Making web services more adaptive

Our work addresses the problem of maintaining the quality of service (QoS) of an orchestration of Web services (WS), which can be affected by exogenous events (i.e., faults). The main challenge in dealing with this problem is that typically the service where a failure is detected is not the one where a fault has occurred: faults have cascade effects on the whole orchestration of services. We have proposed a novel methodology to treat the problem that is not based on Web service (re)composition, but on an adaptive re-execution of the original orchestration. The re-execution process is driven by an orchestrator Manager that takes advantage of an abstract representation of the whole orchestration and may call a diagnostic module to localize the source of the detected failure. It is in charge of deciding the service activities whose results can be reused and may be skipped, and those that must be re-executed.

This year, we have improved the prototype, adding the visualization of the roadmap and the activities that do not have to be reexecuted. This work has been published in ICWS2013 [15] and we are working on a journal paper that will be submitted in 2014.

6.1.2. Scenario patterns for exploring qualitative ecosystems

This work aims at giving means of exploring complex systems, in our case ecosystems, and more recently agrosystems, specifically herd management systems. We proposed to transform environmental questions about future evolution of ecosystems into formalized queries that can be submitted to a simulation model. The system behavior is represented as a discrete event system described by a set of interacting timed automata, the global model corresponding to their composition on shared events. To query the model, we have defined high-level generic query patterns associated to the most usual types of request scenarios. These patterns are then translated into temporal logic formulas. The answer is computed thanks to model-checking techniques that are efficient for analyzing large-scale systems. Five generic patterns have been defined using TCTL (Timed Computation Tree Logic) “WhichStates”, “WhichDate”, “Stability”, “Always”, “Safety”. Three of them have been implemented using the model-checker UPPAAL.

The approach has first been experimented on a marine ecosystem under fishing pressure. The model describes the trophodynamic interactions between fish trophic groups as well as interactions with the fishery activities and with an environmental context. A paper has been previously published in the Environmental Modelling Software Journal [65]. More recently, a similar approach has been experimented on agrosystems, specifically herd management systems, for which a hybrid model has been built using hierarchical timed automata. This later work has been achieved in the context of Yulong Zhao’s PhD thesis [6] and done in collaboration with our colleagues of INRA.

6.1.3. Controler synthesis for dealing with "How to" queries

We extended the approach to deal with "How to" queries. As before, we rely on a qualitative model in the form of timed automata and on model-checking tools to answer queries. We proposed and compared two approaches to answer questions such as "How to avoid a given situation ?"(safety query). The first one exploits controller synthesis and the second one is a "generate and test" approach. We evaluated these two approaches in the context of an application that motivates this work, i.e. the management of a marine ecosystem and the evaluation of fishery management policies. The results have beenpreviously published in [88].

More recently, we used similar methodological tools to analyze in the context of herd management on a catchment. An hybrid model has been built using hierarchical timed automata and scenarios can be simulated and evaluated using the approach presented in the previous paragraph. In this context, the goal is to identify and analyse the best/optimal farming practices in order to reduce nitrate pollution due to livestock effluents. We proposed to use controler synthesis tools and to couple them with machine learning techniques in order to get the best strategies and to put them on easy-to-use form. This work has been made in the context of Yulong Zhao's PhD thesis [6] and in collaboration with our colleagues of INRA (UMR PEGASE).

6.2. Machine learning for model acquisition

Participants: Sid Ahmed Benabderrahmane, Marie-Odile Cordier, Thomas Guyet, Simon Malinowski, René Quiniou.

Model acquisition is an important issue for model-based diagnosis, especially while modeling dynamic systems. We investigate machine learning methods for temporal data recorded by sensors or spatial data resulting from simulation processes. Our main objective is to extract knowledge, especially sequential and temporal patterns or prediction rules, from static or dynamic data (data streams). We are particularly interested in mining temporal patterns with numerical information and in incremental mining from sequences recorded by sensors.

6.2.1. Representing and mining time series

Time series are sequences of numerical values, e.g. recorded by sensors. Since these series can be huge and subject to noise, they are often transformed into sequences of symbols. The best known symbolic transformation method is SAX (Symbolic Aggregate approXimation) [68]. SAX is based on a piecewise constant approximation method that does not take into account the slope of the time series values in successive windows. We have extended the SAX method by adding a symbolic slope information to the SAX symbols. We have experimented our new representation, 1d-SAX, on three mining tasks. In most of these experiments 1d-SAX leads to a better accuracy than SAX [19].

We have also investigated a probabilistic representation of temporal patterns based on the latent Dirichlet allocation model (LDA). Such patterns can approximate the dynamics of a set of similar multivariate time series. We have experimented the method on hydrological flood time series to extract temporal patterns [7]. The extracted patterns were considered relevant and easy to understand by experts of the domain.

6.2.2. Incremental sequential mining

Sequential pattern mining algorithms operating on data streams generally compile a summary of the data seen so far from which they compute the set of actual sequential patterns. We propose another solution where the set of actual sequential patterns are incrementally updated as soon as new data arrive on the input stream. Our work stands in the framework of mining an infinite unique sequence. Our method [60] provides an algorithm that maintains a tree representation (inspired by the PSP algorithm [71]) of frequent sequential patterns and their minimal occurrences [69] in a window that slides along the input data stream. It makes use of two operations: deletion of the itemset at the beginning of the window (obsolete data) and addition of an itemset at the end of the window (new data). The experiments were conducted on simulated data and on real data of instantaneous power consumption. The results show that our incremental algorithm significantly improves the computation time compared to a non-incremental approach [61].

Recently, we have worked on the adaptation of our algorithm to closed sequential patterns. A closed pattern is a local maximal pattern such there exists no extension of this pattern having the same support. Closed patterns are known to provide a condensed represention of the solution patterns and lead to more efficient algorithms without losing information or completeness on extracted patterns. The tree of closed-patterns is less deep than the pattern-tree but the transformations of the tree by addition or deletion of items are more complex. The algorithm is under evaluation. We plan to submit a paper in 2014.

6.2.3. Multiscale segmentation of satellite image time series

Satellite images allow the acquisition of large-scale ground vegetation. Images are available along several years with a high acquisition frequency (1 image every two weeks). Such data are called satellite image time series (SITS). In [58] , we presented a method to segment an image through the characterization of the evolution of a vegetation index (NDVI) on two scales: annual and multi-year. The main issue of this approach was the required computation resources (time and memory). We first propose to adapt image segmentation algorithm to SITS. Segmented images reduces the number of time series to analyze and the computation time. We secondly applied 1D-SAX to reduce data dimensionality [20]. We evaluated this approach on the supervised classification of large SITS of Senegal and we showed that 1D-SAX approaches the classification results of time series while significantly reducing the required memory storage of the images.

6.2.4. Analysis of landscape based on spatial patterns

Researchers in agro-environment need a great variety of landscapes to test the agro-ecological models of their scientific hypotheses. Real landscapes are difficult to acquire and do not enable the agronomist to test all their hypothesis. Working with simulated landscapes is then an alternative to get a sufficient variety of experimental data. Our objective is to develop an original scheme to generate realistic landscapes. This approach is based on a spatial representation of landscapes by a graph expressing the spatial relationships between the agricultural parcels (as well as the roads, the rivers, the buildings, etc.), of a specific geographic area. We extract spatial patterns from a real geographic area and we use these patterns to generate new realistic landscapes. Using patterns preserves the interface properties between parcels.

We have begun the exploration of graph mining techniques, such as gSPAN [87], to discover the relevant spatial patterns present in a spatial-graph. But the graph-mining techniques are very time-consuming in comparison to sequence mining.

This year, we would like to test if using a path instead of a graph would be a faithful representation of the spatial organization of the landscape. In [17], we compare the potential expressivity of graphs and Hilbert-Peano curves [66] to characterize an agricultural landscape. The results show that mining frequent patterns in Hilbert-Peano curves would be as discriminant as mining frequent patterns in graphs.

The perception of the environment is an important dimension of the landscape we live in. One of our objectives is to study the relationships between the landscape patterns and their perception. We cope with this dimension by analysing the textual content of ‘atlas du paysage" (landscape atlas), that are produce by each french administrative regions. This year we worked on the construction of an ontology of landscape perception [21].

6.2.5. Subdimensional clustering for fast similarity search over time series data. Application to Information retrieval tasks

Information retrieval and similarity search tasks in time series databases remains a challenge that require to discover relevant pattern-sequences that are recurrent over the overall time series sequences, and to find temporal associations among these frequently occurring patterns. Previous work on information retrieval and similarity search in time series has been performed in different contexts such as diagnosis or failure detection of industrial materials. In whole query matching, a time series given as query is entirely compared to every time series of a database. The series should have same length, and a similarity measure is used to retrieve either a most similar time series or the top-k ranked time series. However, theses methods suffer from a lack of flexibility of the used similarity measures, a lack of scalability of the representation model, and a penalizing runtime to retrieve the information. Moreover, in some real world applications, one can be interested in retrieving specific interesting subsequences that are frequently present at different instants.

Motivated by these observations, we have designed a framework tackling the query by content problem on time series data, ensuring (i) fast response time, (ii) multi-level information representation, and (iii) representing temporal associations between extracted patterns. During the preparation step, all the multi-valued time series present in the database are transformed into a multi-resolution symbolic representation thus ensuring a lower dimensionality. Then, to accelerate and enhance the similarity search and the retrieval over the database, our model creates an index over recurrent patterns in the time series collection. These patterns can be generated by different techniques. Finally, the extracted patterns are grouped by clustering and the resulting clusters are indexed in a table within their centroids. A paper presenting the preliminary results is under submission to an international journal.

6.2.6. Knowledge Extraction from Heterogeneous Data

Recently, mining microarrays data has became a big challenge due to the growing sources of available data. We are using machine learning methods such as clustering, dimensionality reduction, association rules discovery on transcriptomic data, by combining a domain ontology as source of knowledge, in order to supervise the KDD process. Our objectives concern the identification of genes that could participate in the development of tumors. A two-way classification method was proposed, combining genes expression levels, represented as numerical data, and Gene Ontology (GO) annotations as symbolic data. The hopeful results obtained with genes clustering, through GO annotations, are an encouraging track to predict transcriptional regulatory networks, and for refining the existing sets of genes [11], [12].
We also introduced a new method for extracting enriched biological functions from transcriptomic databases using an integrative bi-classication approach. The initial gene datasets are firstly represented as a formal context (objects attributes), where objects are genes, and attributes are their expression profiles and complementary information of different knowledge bases. Formal Concept Analysis (FCA) is applied for extracting formal concepts regrouping genes having similar transcriptomic profiles and functional behaviors. An enrichment analysis is then performed in order to identify the relevant formal concepts from the generated Galois lattice, and to extract biological functions that could participate in the proliferation of cancers. Preliminary results seem very promising, and could help experts during the identification of degenerated biological functions [13].

6.3. Decision aiding with models and simulation data

Participants: Louis Bonneau de Beaufort, Tassadit Bouadi, Marie-Odile Cordier, Véronique Masson, René Quiniou.

Models can be very useful for decision aiding as they can be used to play different plausible scenarios for generating the data representing future states of the modeled process. However, the volume of simulation data may be huge. Thus, efficient tools must be investigated in order to store the simulation data, to focus on relevant parts of the data and to extract interesting knowledge from these data.

6.3.1. A datawarehouse for simulation data

The ACASSYA project 8.2.1 aims at providing experts or stakeholders or farmers with a tool to evaluate the impact of agricultural practices on water quality. As the simulations of the deep model TNT2 are time-consuming and generate huge data, we have proposed to store these simulation results in a datawarehouse and to extract relevant information, such as prediction rules, from the stored data. We have devised a general architecture for agro-environmental data on top of the framework Pentaho.

This year we have been working on the efficient computation of OLAP queries related to realistic scenarios proposed by experts in the domain. Precisely, we have devised indexing schemes to access the data in the OLAP cube. We have also worked on the visualization by a GIS (Geographical Information System) of the query results on maps of the geographical area under interest. A paper have been submitted to the COMPAG Journal. This work is detailed in Tassadit Bouadi's thesis [5].

6.3.2. Efficient computation of skyline queries in an interactive context

Skyline queries retrieve from a database the objects that maximize some criteria, related to user preferences for example, or objects that are the best compromises satisfying these criteria. When data are in huge volumes, such objects may shed light on interesting parts of the dataset. However, computing the skylines (i.e. retrieving the skyline points) may be time consuming because of many dominance tests. This is, especially the case in an interactive setting such as querying a data cube in the context of a datawarehouse.

We have worked at improving the formal setting of the partial materialization of skyline queries when dynamic preferences are refined online by the user. We have explicited which parts of the skyline evolve (which point are added or removed) when a new dimension is introduced in the computation. This led to an efficient incremental method for the online computation of the skyline corresponding to new user preferences [46]. An extended version of this paper is published in Journal "Transactions on Large Scale Data and Knowledge Centered Systems" (TLDKS) [8] and in Tassadit Bouadi's thesis [5].

6.3.3. Hierarchical skylines

Conventional skyline queries retrieve the skyline points in a context of dimensions with a single hierarchical level. However, in some applications with multidimensional and hierarchical data structure (e.g. data warehouses), skyline points may be associated with dimensions having multiple hierarchical levels. Thus, we have proposed an efficient approach reproducing the effect of the OLAP operators "drill-down" and "roll-up" on the computation of skyline queries. It allows the user to navigate along the dimensions hierarchies (i.e. specialize / generalize) while ensuring an online calculation of the associated skyline. The method is described in Tassadit Bouadi's thesis [5]. A paper describing this contribution is currently under submission to the "Very Large Data Bases (VLDB 2014)" conference.

6.3.4. Modeling influence propagation by Bayesian causal maps

The goal of this project is modeling shellfish fishing to assess the impact of management pollution scenarios on the *Rade de Brest*. Cognitive maps were built from interviews with fishermen. To represent and reason about these cognitive maps, we propose to use Bayesian Causal Maps making use of fishermen knowledge, particularly to perform influence propagation [82].

However, this model does not take into account the variety of influences asserted by the fishermen, but only the "mean" causal map. A report describing the project is available [28]. An approach that could combine individual knowledge with belief functions in the way of Philippe Smets's Transferable Belief Model [83] has been proposed. A report describing the projectis available [28].

This work is done in the framework of the RADE2BREST project, involving Agrocampus Ouest and CNRS (GEOMER/LETG), funded by "Ministère de l'Ecologie" (This project is not mentioned in section 8.2 because DREAM is not an official partner of this project.).

6.3.5. Recommending actions from classification rules

In the framework of the SACADEAU project, a paper dedicated to building actions from classification rules has been published in the KAIS Journal [9]. Our goal is to burden of analysing a large set of classification rules when the user is confronted to an "unsatisfactory situation" and needs help to decide about the appropriate actions to remedy to this situation. The method consists in comparing the situation to a set of classification rules. For this purpose, we propose DAKAR, a new framework for learning action recommendations dealing with complex notion of feasibility and quality of actions.

Sacadeau-Software, which is the decision support tool implemented with F. Ployette (former Inria engineer in the EPI Dream, now retired) in the SACADEAU project, has been published in the RIA Journal [10]. Sacadeau-Software allows to run simulations throughout a watershed and obtain the transfer rate of pollution through the catchment. Classification rules, characterizing the sub-parts of the watershed with pollution and the sub-parts without pollution, are automatically learned from the simulations. A visualization tool enables to relate the learned rules to the examples characterized by these rules. Finally, a user can select a situation of pollution

and the action recommendation tool analyses the learned rules and proposes actions that improve this situation of pollution.

6.4. Diagnostic, causal reasoning and argumentation

Participants: Philippe Besnard, Marie-Odile Cordier, Yves Moinard.

Stemming on [38], [39], [40], [41], [42], we have designed an inference system based on causal statements. This is related to diagnosis (observed symptoms explained by faults). The aim is to produce possible explanations for some observed facts. Previously existing proposals were ad-hoc or, as in [45], [57], they were too close to standard logic to make a satisfactory diagnosis. A key issue for this kind of work is to distinguish logical implication from causal links and from ontological links. This is done by introducing a simple causal operator, and an *is-a* hierarchy. These two operators are added to a restricted first order logic of the Datalog kind (no function symbols). Then, our system produces elementary *explanations* for some set of observed facts. Each explanation links some facts to the considered observation, together with a set of atoms called the *justifications*: The observation is explained from these facts, provided the justifications are possible (not contradicted by the available data). This formalism has been translated into answer set programming [72], [73]). It is able to deal with complex problems such as finding explanations for the hurricane Xynthia (2010, February 28). In such situations, there are many data and many possible elementary explanations can be examined. This involves an extension of our formalism, in order to deal with more complex chains of causations and *is-A* links. Our formalism makes precise what all these possible explanations are. Then, in order to deal with so many possible complex explanations, we integrate this causal formalism into an argumentation framework. Logic-based formalizations of argumentation [43] take pros and cons for some conclusion into account. These formalizations assume a set of formulae and then exhaustively lay out arguments and counterarguments. This involves providing an initiating argument for the inference and then providing undercuts to this argument, and then undercuts to undercuts. So here our causal formalism provides a (rather large) set of explanations, and the argumentation part allows to select the best ones, under various criteria [22], [14].

Then, since answer set programming can easily deal with logical formalisms, the argumentation part will be incorporated into our already existing answers set programming translation of the causal formalism. Regarding this field of knowledge representation and reasoning, and more generally, artificial intelligence, we have participated to several chapters in the to be published "Panorama de l'intelligence artificielle. Ses bases méthodologiques, ses développements" [27], [26], [23], [24].

DYLISS Project-Team

6. New Results

6.1. Data integration

Participants: Jacques Nicolas, Andres Aravena, Charles Bettembourg, Jérémie Bourdon, Jeanne Cambefort, Guillaume Collet, Olivier Dameron, Damien Eveillard, Julie Laniau, Sylvain Prigent, Anne Siegel, Sven Thiele, Valentin Wucher.

Metabolic network reconstruction: combinatorial gap-filling method We introduced an exhaustive gap-filling procedure on the first metabolic network for a macroalgea (Ectocarpus Siliculosus). As this species is a non benchmark model, this issue is related to hard combinatorial optimization problems. To that matter, we took advantages of the latest improvement of Answer Set Programming solvers (combination of clasp and unclasp) and introduced a new model of the network expansion problem. [*G. Collet, D. Eveillard, S. Prigent, A. Siegel, S. Thiele*] [27]

Identification of functional gene units in non benchmark models We introduced the concept of "shortest genome segments" (SGS) to detect functional units on exotic species, such as extremophiles, that are by nature unrefined. They correspond to genome portion which contain a large density of genes coding for enzymes which regulate successive reactions of metabolic pathways. There identification is a hard optimization combinatorial problem. We relied on the declarative modeling power of answer set programming (ASP) to encode the identification of shortest genome segments and prove that SGS are stable in (i) computational time and (ii) ability to predict functional units when one deteriorates the biological knowledge [*D. Eveillard, A. Siegel, S. Thiele*] [26]

Refinement of regulatory network from genomic, expression data and functional unit data We integrated heterogeneous information from two types of network predictions to determine a causal explanation for the observed gene co-expression. We modeled this integration as a combinatorial optimization problem. We demonstrated that this problem belongs to the NP-hard complexity class. We proposed an heuristic approach to have an approximate solution in a practical execution time. Our evaluation showed that the E.coli regulatory network resulting from the application of this method has higher accuracy than the putative one built with traditional tools. Applications to the mining bacterium *Acidithiobacillus ferrooxidans* allowed analyzing the relevance of central regulators. [*A. Aravena, D. Eveillard, A. Siegel*] [23], [13] [Thesis]

Reconstruction of a protein interaction network for archaebacteria To gain insights into genomic maintenance processes in hyperthermophilic archaea, a protein-interaction network centered on informational processes of *Pyrococcus abyssi* was generated by affinity purification coupled with mass spectrometry. We have proposed a graph theoretic analysis of this network including statistical (e.g. clusterisation coefficients) and topological aspects (bicluster analysis, search of a maximal interaction skeleton), which helps network interpretation in terms of formation of complexes or interaction dynamics. [*J. Nicolas*] [20] [Online publication]

Knowledge evolution in ontologies We studied the impact of an ontology evolution on its structural complexity. As a case study we used sixty monthly releases of the Gene Ontology and its three independent branches i.e. biological processes (BP), cellular components (CC) and molecular functions (MF). For each release, we measured complexity by computing metrics related to the size, the nodes connectivity and the hierarchical structure. We showed that the variation of the number of classes and relations in an ontology does not provide enough information about the evolution of its complexity. However, connectivity and hierarchy-related metrics revealed different patterns of values as well as of evolution for the three branches of the Gene Ontology [*O. Dameron, C. Bettembourg*] [17], [14] [Online publication][Thesis]

Treatment process representation for breast cancer patients. The general cancer registry of Poitou-Charentes developed a multiple source information system covering diseases, anatomical structures and cytopathology. We proposed an algorithm for representing and analyzing the patient's treatment process. An expert compared the original data with our representation and computed a score of dissimilarity. The results showed that an integrated information system can successfully analyze the data to determine whether they comply with the guidelines [*O. Dameron*] [31].

AphidAtlas project We began a collaboration with the AphidAtlas project for defining the structure of an ontology of aphids anatomy and development [*O. Dameron*] [30].

6.2. Asymptotic dynamics

Participants: Anne Siegel, Oumarou Abdou-Arbi, Geoffroy Andrieux, Jérémie Bourdon, Jeanne Cambefort, Damien Eveillard, Michel Le Borgne, Vincent Picard, Sven Thiele, Santiago Videla.

Learning families of boolean signaling networks We propose the use of ASP to explore the space of feasible logic models of a signaling network. To that matter, we exhaustively enumerate the set of sub-optimal boolean logical models which are compatible with both the topology of a knowledge-based influence graph and the observed response of the system to several perturbations (phosphorylation datasets). We illustrate the importance of characterizing such a family of models in a global and exhaustive manner by revisiting a model of pro-growth and inflammatory pathways in human liver cells and studying the variability with the set of compatible models. [*A. Siegel, S. Thiele, S. Videla*] [18] [Online publication]

Control the steady-state response of qualitative signaling networks: intervention sets The minimal intervention set problem roughly consists in identifying the perturbation that can be undergone over a signaling network to predict a fixed expected behavior. We have provided a precise characterization of the minimal intervention set problem relying on three-valued logic and fixpoint semantics. We address this problem within ASP and using real-world biological benchmarks we show that it greatly outperforms previous work using dedicated algorithms. [*A. Siegel, S. Videla*] [19] [Online publication]

Reachability in dynamical signaling networks: cut sets In the scope of discrete finite-state models of interacting components, we present a novel algorithm for identifying sets of local states of components whose activity is necessary for the reachability of a given local state. Those sets are referred to as cut sets; they provide potential therapeutic targets that are proven to prevent molecules of interest to become active, up to the correctness of the model. Our method is based on the so-called Graph of Local Causality and form an under-approximation of the complete minimal cut sets of the dynamics. It makes tractable the formal analysis of very large scale networks. [*G. Andrieux, M. Le Borgne*] [28], [12] [Online publication][Thesis]

Exploring metabolism flexibility through quantitative study of precursor sets for system outputs We extended a Flux-Balanced-Analysis approach to quantify the precursor composition of each system output and to discuss the biological relevance of a set of flux in a given metabolic network. The composition is called contribution of inputs over outputs [AIO]. In order to further investigate metabolic network flexibility, we have proposed an efficient local search algorithm computing the extremal values of AIO coefficients. This approach enables to discriminate diets without making any assumption on the internal behaviour of the system. [*J. Bourdon, O. Abdou-Arbi, A. Siegel*] [15], [11] [Thesis]

6.3. Sequence annotation

Participants: François Coste, Aymeric Antoine-Lorquin, Catherine Belleannée, Guillaume Collet, Gaëlle Garet, Clovis Galiez, Laurent Miclet, Olivier Quenez, Jacques Nicolas, Valentin Wucher.

Refinement of mi-RNA regulation network thanks to concept analysis MicroRNAs (miRNAs) are small RNA molecules that bind messenger RNAs (mRNAs) to silence their expression. To improve the discrimination between true and false interactions during their prediction, we defined a repair process based on the hypothesis that the true graph is formed by interaction modules represented by formal concepts, i.e. set of miRNAs having the same regulation profile. To validate our hypothesis and method, we have extracted parameters from a biological miRNA/mRNA network and used them to build random networks. Each repaired

network can be evaluated with a score balancing the number of edge changes and the conceptual adequacy in the spirit of the minimum description length principle. [*J. Nicolas, V. Wucher*] [32]

Analogical proportions and the factorization of information in distributive lattices. We have conducted theoretical studies to elucidate whether formal concept lattices can have properties that could be used in further studies. In this direction, analogical proportions are statements involving four entities, of the form 'A is to B as C is to D'. They play an important role in analogical reasoning. They have been formalized in both a propositional logic setting and an algebraic setting. We define and study analogical proportions in the general setting of lattices, and more particularly of distributive lattices. We discussed the decomposition of analogical proportions in canonical proportions as well as the resolution of analogical proportion equations, and illustrate especially on the case of Boolean lattices, which reflects the logical modeling. [*L. Miclet*] [24], [29]

Bioinformatics and Artificial Intelligence In this book chapter, we introduce the main objects studied in Bioinformatics at different levels (the macromolecules, their interactions as well as the knowledge formalization or extraction) and present meanwhile a survey of the contribution and influence of Artificial Intelligence to this research field on related key tasks (gene prediction, functional annotation, structure prediction, transcriptomics analysis, network acquisition and analysis, knowledge integration and formalization, information retrieval and extraction from documents, ...). [*F. Coste*] [33]

Genome studies: fast assembly and SNP identification This work is a follow-up of collaborations with the GenScale team and the GenOuest platform. We reported the first identification of a set of SNPs isolated from the genome of *I. ricinus* - an important vector of pathogens in Europe, by applying a reduction of genomic complexity, pyrosequencing and new bioinformatics tools[21] [Online publication]. We also contributed to show that the genome assembly program MINIA is successfully able to assemble a 100 Mbp genome on a very low-end, low-power system with 512 MB RAM and a 32 GB flash drive such as a Raspberry Pi. [*G. Collet, O. Quenez*] [34][Online publication]

ESPRESSO Project-Team

6. New Results

6.1. A pivot in between synchrony and asynchrony

Participants: Thierry Gautier, Paul Le Guernic, Jean-Pierre Talpin.

Our time modeling framework requires a pivot specification paradigm to materialise a spectrum of models of computation and communication ranging from synchrony to asynchrony, from software to hardware, and accommodate with (abstractions of) software behaviors (software, functional blocks, tasks) and requirements (temporal properties, contracts, regular expressions) through logical, periodic, multi-periodic or affine time. We aim at developing a framework comprising dataflow networks (communications) and synchronous automata (computations) controlled by synthesised wrapper enforcing abstractions of specified constraints from the software viewpoint (timing requirements).

Relations between Kahn networks and classes of synchronous dataflow graphs (SDF) as well as synchronous languages have been studied in the past (e.g. Lustre), yet never in the full generality of relating the domain-theoretic model of Kahn networks to it most general synchronous incarnation (one that at least allows to express several clock domains) [17]. We are currently elaborating such a model to characterise morphisms between untimed asynchronous networks and multi-clocked, synchronous, dataflow networks. In this prospect, we developed the first constructive operational semantics of Signal [21], which opens to further investigations on its full abstraction relation with a denotational characterisation using Kahn networks over a polychronous domain.

6.2. New functionalities of Polychrony

Participants: Loïc Besnard, Thierry Gautier, Paul Le Guernic.

We have developed and integrated in the Signal toolbox some clock computations useful for optimizations: *assignment clocks* and *utility clocks*. These information may be used to reduce the frequency of the computations and the communications for distributed code generation.

Assignment clock. A given signal is supposed to be computed at the instants of its clock, defined by the clock of the expression of its definition. For a signal x, the expression of its definition can always be rewritten as $x := (E_1 \texttt{ when } h_1) \texttt{ default } ... \texttt{ default } (E_{n-1} \texttt{ when } h_{n-1}) \texttt{ default } (x \$ \texttt{ when } k)$. If we assume that the signal keeps, between two consecutive instants, the last computed value, the assignment of $(x \$)$ to x is unnecessary. Then, the assignment clock is then defined by $h_1 \hat{+} ... \hat{+} h_{n-1}$, smaller than the clock of x defined by $(h_1 \hat{+} ... \hat{+} h_{n-1}) \hat{+} k$.

Utility clock. The utility clock defines the instants at which a signal is necessary. For a signal x, the utility clock, $hu(x)$ is defined by:

- the clock of x if x is an input, an output, a memory, or if it is used to define an undersampling clock ($\texttt{when } f(x)$);
- otherwise, it is defined, for $x \rightarrow y_1 \texttt{ when } h_1$, ..., $x \rightarrow y_n \texttt{ when } h_n$, by $\sum_{i=1,n} (hu(y_i) \hat{*} h_i)$.

If we rewrite the Signal program by sampling the signals (except for inputs/outputs) by their utility clock, the new Signal program is equivalent to the previous one, with respect to its behavior with the external world. Note that the utility clock can be used only when this transformation does not introduce cycles in the graph.

6.3. Formal Verification of Synchronous Dataflow Program Transformations Toward Certified Compilers

Participants: Van-Chan Ngo, Jean-Pierre Talpin, Thierry Gautier, Paul Le Guernic, Loïc Besnard.

Translation validation [49], [48] is a technique that attempts to verify that program transformations preserve the program semantics. A compiler generally involves several phases during its compilation process. For instance, the Signal compiler [2], [8], in its first two phases, *calculates the clock information*, makes *Boolean abstraction*, and makes *static scheduling*. The final phase is the executable code generation. One can try to prove globally that the input program and its final transformed program have the same semantics. However, we believe that a better approach consists in separating the concerns and proving for each phase the preservation of different kinds of semantic properties. In the case of the Signal compiler, the preservation of the semantics can be decomposed into the preservation of clock semantics, data dependence, and value-equivalence of variables.

Translation Validation for Clock and SDGs Transformations. This work focuses on proving the preservation of clock semantics in the first two phases of the Signal compiler. In order to do that we encode the clock semantics and data dependence as *clock models* and *synchronous dependence graphs* (SDGs). Then we show that a transformation is correct if and only if there exist *refinements* between clock models, and between SDGs, written as $\Phi(P_2) \sqsubseteq_{clk} \Phi(P_1)$ and $SDG(P_2) \sqsubseteq_{dep} SDG(P_1)$ [15]. We delegate the checking of the preservation to a SMT-solver [38], [54].

Translation Validation of Polychronous Dataflow Specifications: from Signal to C using Synchronous Dataflow Value-Graphs. In this work, we build a validator for the synchronous dataflow compiler of Signal. This validator tries to match the value-graph [53] of each output of the original program and its transformed counterpart. That ensures that every output of the original program and its counterpart in the transformed program have the same value whenever they are present. Our validator does not require any instrumentation and modification of the compiler, nor any rewriting of the source program.

The Signal program and its generated C program have been represented in the same shared synchronous dataflow value-graph (SDVG), in which the nodes for the same structures (variables, constants, operators) have been shared. For instance, the values of input signals and their corresponding variables in the generated C code are represented by the same nodes in the shared graph. Then, the shared graph is transformed following *predefined rules* to show that all output signal values in the Signal program and their counterparts in the generated C code are rooted at the same subgraph.

Consider the following process, where $IR(P)$ is the compiled code of the program P and TV(SDVG(P,IR(P))) is *true* when all output signal values in P and their counterparts in $IR(P)$ are the same:
if (Cp(P) is Error) then output Error; else
if ($(\Phi$(IR(P)) $\sqsubseteq_{clk}$ Φ(P)) and (SDG(IR(P)) $\sqsubseteq_{dep}$ SDG(P)) and (TV(SDVG(P,IR(P))))) then output IR(P); else output Error.
This will provide formal guarantee as strong as that provided by a formally certified compiler w.r.t. the clock semantics and the data dependence in case the validator is certified formally.

Implementation and Experiments. At a high level, our tool *SigCert* [47] developed in OCaml checks the correctness of the compilation of the Polychrony Signal compiler w.r.t clock semantics, data dependence, and value-equivalence as shown in Figure 8 .

6.4. Exploring system architectures in AADL via Polychrony and SynDEx

Participants: Huafeng Yu, Loïc Besnard, Thierry Gautier, Jean-Pierre Talpin, Paul Le Guernic.

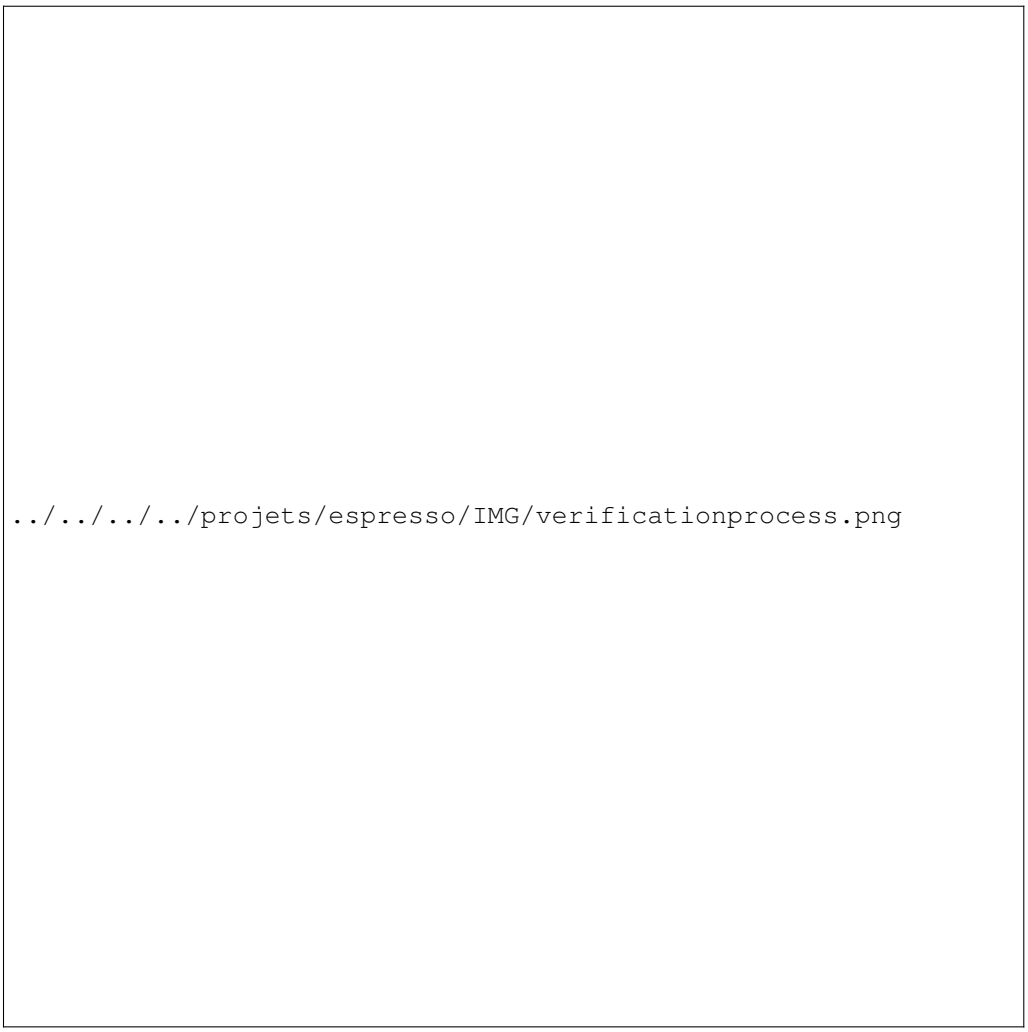


Figure 8. An overview of our integration within Polychrony toolset.

Architecture analysis & design language (AADL) has been increasingly adopted in the design of embedded systems, and corresponding scheduling and formal verification have been well studied. However, little work takes code distribution and architecture exploration into account, particularly considering clock constraints, for distributed multi-processor systems. Our approach [20], [16], [17] handles these concerns within the toolchain AADL-Polychrony-SynDEx. First, in order to avoid semantic ambiguities of AADL, the polychronous/multiclock semantics of AADL, based on a polychronous model of computation, is considered. Clock synthesis is then carried out in Polychrony, which bridges the gap between the polychronous semantics and the synchronous semantics of SynDEx [42]. The same timing semantics is always preserved in order to ensure the correctness of the transformations between different formalisms. Code distribution and corresponding scheduling is carried out on the obtained SynDEx model in the last step, which enables the exploration of architectures originally specified in AADL. Our contribution provides a fast yet efficient architecture exploration approach for the design of distributed real-time and embedded systems. The approach has been illustrated using an avionic case study.

6.5. A synchronous annex for the AADL

Participants: Loïc Besnard, Thierry Gautier, Paul Le Guernic, Jean-Pierre Talpin.

We propose a synchronous timing annex for the SAE standard AADL. Our approach consists of building a synchronous model of computation and communication that best fits the semantics and expressive capability of the AADL and its behavioral annex and yet requires little to know (syntactic) extension to it, i.e. to identify a synchronous core of the AADL (which prerequisites a formal definition of synchrony at hand) and define a formal design methodology to use the AADL in a way that supports formal analysis, verification and synthesis.

Our approach first identifies the core AADL concepts from which time events can be described. Then, is considers the behavior annex (BA) as the mean to model synchronous signals and traces through automata. Finally, we consider elements of the constraint annex to reason about abstractions of these signals and traces by clocks and relations among them. To support the formal presentation of these elements, we define a model of automata that comprises a transition system to express explicit transitions and constraints, in the form of a boolean formula on time, to implicitly constraint its behavior. The implementation of such an automaton amounts to composing its explicit transition system with that of the controller synthesised from its specified constraints.

6.6. Ongoing activities and results for integration of Polychrony with the P toolset

Participants: Christophe Junke, Loïc Besnard, Thierry Gautier, Paul Le Guernic, Jean-Pierre Talpin.

Current state of P. The P language is still under definition, notably for the software/hardware architectural description of systems. In late october 2013, technical partners (headed by Adacore) released the first beta version of the toolset. The main activites of the ESPRESSO team can be splitted in analysis and development activities:

- The analysis activites consisted in understanding what tasks shall be performed ultimately by the P toolset w.r.t. code generation and architecture, and how Polychrony could be used in the proposed workflow.
- The development activities consisted in introducing a modified block sequencing algorithm in P and starting the development of the P to Signal converter.

Co-modeling in P. First, P should import functional behavior from Simulink, Stateflow and UML class, activity and state machine diagrams. Those represent a strictly sequential semantics: “the code generated from functional behaviour language will be strictly sequential and void of tasking features” (P specification [1]).

[1] https://forge.open-do.org/plugins/moinmoin/p/

Second, imported architectural description languages are likely to be SysML, MARTE and AADL, which present concurrent semantics. Hence, "the code generated from architectural description languages may include concurrent semantics (thread, shared resources...)" (*ibid*). However, code generation from architectural description languages will consist of invocations to an underlying real-time API. The current target of code generation is the APEX ARINC-653 API, which provides real-time services like inter/intra-partition communication channels as well as task scheduling. Real-time properties of imported architectural elements, like task periods and scheduling policy, are used to configure those services.

Code distribution. Code generation should be able to distribute the functional blocks among architectural elements (processors/threads and buses/queues).

Polychrony offers a way to distribute Signal processes among different locations [31]. In general, such code distribution may lead to the synthesis of new input and output ports: when expressing synchronous communication with asynchronous protocols, some clock information might need to be added to resynchronize data-flows. Moreover, the computation model of Signal allows to order asynchronous read and write operations to avoid communication deadlock. The extended input/output interfaces of blocks could be reimported back to P in order to ensure the correctness of code distribution.

It appears however that the subset of Simulink that is imported in P, and the execution model of P functional models that is enforced by the P compiler, can be viewed as a composition of endochronous multi-rate nodes (all inputs of a node are computed before all of its outputs; this avoids deadlock problems when composing nodes). This model ends up being similar to a Lustre model of computation, where code distribution can be performed without adding communication flows and where read/write operations can be setup in a general way without introducing deadlocks [41].

Despite the above observations, it might be possible to extend the input/output interfaces of existing P models thanks to Polychrony. One approach is to ensure that block dependencies between Simulink blocks are effectively respected after code distribution. Indeed, functional blocks can be partially ordered thanks to user-defined priorities. If other partners see an interest with this approach, it could be possible to establish communication links between ordered blocks, so that the global execution order of blocks in a distributed setting is the same as the one modeled originally in the simulation environment.

Model clustering. Alternatively, it would be interesting from a Signal point of view to loosen the synchronization assumptions made by both Simulink and P so that only algebraic dependencies are taken into account (e.g. interpret all Simulink subsystems as virtual, ignore all non-strictly required dependencies...), while respecting clock constraints (e.g. sample time, controlled and enabled blocks...). In that case, the Signal compiler could perform code distribution for simulation purposes, or simply to provide another compilation scheme for P. Another step could be to apply an automatic code distribution mechanism into so-called *clusters*, and export those clusters back to P as architectural elements. The resulting P model would end-up being having possibly more tasks/threads and smaller functional blocks, which might be interesting. Those design decisions are still under consideration and must be discussed with other partners.

From P to Signal. The development activities in the P project currently consist in adding a P to Signal translator. It is being developed as a backend of the P toolset, which provides a number of facilities to access and perform computations on P models. The current prototype must be completed and refined according to what are the actual needs in the project, but can already be tested with input models.

In order to validate the approach, the existing test models of the P projects are all checked with this exporter (over two hundreds small models, a couple of big ones). The resulting SSME files are then converted to Signal files: this step required to generate a command-line version of the Eclispe Polychrony product, as well a a batch converter from SSME to Signal (this converter is integrated in the Polychrony environment). In addition to convert SSME files to Signal, this converter also validates the model against Ecore constraints. Finally, the resulting Signal files are compiled with the original `C++` Signal compiler to check typing and clock relationships (those tests are not performed at the SSME level). The resulting test toolchain gives useful feedbacks for the iterative development of the translator.

Partial orders in P. The exporter also needs to export block dependencies from functional models. Since Polychrony is also able to infer a total order while taking into account code distribution, it was not satisfactory to export the existing total order computed by the P toolset: it is more sensible to export the subset that is strictly necessary (or desired). In agreement with technical partners, we modified the existing sequencer so that it could be parameterized with block ordering criteria (for example, we might want to take into account dataflow dependencies as well as user-defined priority in Polychrony, but nothing more). The outcome is a single package responsible for computing partial and total orders inside the P toolset. This prevents other tools, like the P to Signal exporter, to compute partial order by themselves.

The implementation of the sequencer is based on a dependency matrix that helps computing the transitive closure of dependencies (to quickly check whether two blocks are dependent on each other) while keeping track of their transitive reduction (in order to export only the minimal set of relationships). Now that the first version of the P toolset is released, the sequencer will hopefully be integrated in the P toolset.

6.7. Real-Time Scheduling of Dataflow Graphs

Participants: Adnan Bouakaz, Jean-Pierre Talpin.

The ever-increasing functional and nonfunctional requirements in real-time safety-critical embedded systems call for new design flows that solve the specification, validation, and synthesis problems. Ensuring key properties, such as functional determinism and temporal predictability, has been the main objective of many embedded system design models. Dataflow models of computation (such as KPN [44], SDF [46], CSDF [34], etc.) are widely used to model stream-based embedded systems due to their inherent functional determinism. Since the introduction of the (C)SDF model, a considerable effort has been made to solve the static-periodic scheduling problem [28]. Ensuring boundedness and liveness is the essence of the proposed algorithms in addition to optimizing some nonfunctional performance metrics (e.g. buffer minimization, throughput maximization, etc.). However, nowadays real-time embedded systems are so complex that real-time operating systems are used to manage hardware resources and host real-time tasks. Most of real-time operating systems rely on priority-driven scheduling algorithms [51], [37] (e.g. RM, EDF, etc.) instead of static schedules which are inflexible and difficult to maintain. Our work [12], [18], [19] [35] addresses the real-time scheduling problem of dataflow graph specifications; i.e., transformation of the dataflow specification to a set of independent real-time tasks w.r.t. a given priority-driven scheduling policy such that the following properties are satisfied: (1) channels are bounded and overflow/underflow-free; (2) the task set is schedulable on a given uniprocessor (or multiprocessor) architecture. This problem requires the synthesis of scheduling parameters (e.g. periods, priorities, processor allocation, etc.) and channel capacities. Furthermore, our work considers two performance optimization problems: buffer minimization and throughput maximization.

6.8. Structure-Preserved Distribution of Synchronous Programs

Participants: Ke Sun, Loïc Besnard, Thierry Gautier, Paul Le Guernic, Jean-Pierre Talpin.

We propose an automatically structure-preserved distribution method, which is based on synchronous guarded actions [50] and component calls in an intermediate representation [36]. The guarded actions describe the local behavior. The component calls preserve the modular structure information of synchronous programs. Using this method, the designer can naturally blend the distribution design into the whole system design procedure: following the modular structure, a globally asynchronous locally synchronous (GALS) network over distributed nodes can be automatically constructed. Each node corresponds to a component and contains:

- a computing element, computing (as sender) or reacting to (as receiver) scheduling commands;
- a controlling element, called adaptor, adjusting the asynchronous communication between nodes.

The computing element focuses on the functional behaviors (i.e., value computation) in synchronous runs, which can be perfectly described by synchronous guarded actions. On the other hand, the controlling element mainly considers the temporal constraints (i.e., clock relation) under asynchronous communications. Guarded actions are not suitable for specifying clock relations, then we use polychronous specifications [8] to define the inter-node communications.

A perspective for future work would be the structure-preserved distribution of synchronous programs with multi-interaction. Multiple interactions in one logical instant are desynchronized and projected onto finer grained instants. Owing to this extension, it would provide more convenience for the designer to express the parallelism among components.

FLUMINANCE Project-Team

6. New Results

6.1. Fluid motion estimation

6.1.1. Stochastic uncertainty models for motion estimation

Participants: Sébastien Beyou, Etienne Mémin, Emmanuel Saunier.

In this study we have proposed a stochastic formulation of the brightness consistency used principally in motion estimation problems. In this formalization the image luminance is modeled as a continuous function transported by a flow known only up to some uncertainties. Stochastic calculus then enables to built conservation principles which take into account the motion uncertainties. These uncertainties defined either from isotropic or anisotropic models can be estimated jointly to the motion estimates. Such a formulation besides providing estimates of the velocity field and of its associated uncertainties allows us to naturally define a linear multiresolution scale-space framework. The corresponding estimator, implemented within a local least squares approach, has shown to improve significantly the results of the corresponding deterministic estimator (Lucas and Kanade estimator). This fast local motion estimator provides results that are of the same order of accuracy than state-of-the-art dense fluid flow motion estimator for particle images. The uncertainties estimated supply a useful piece of information in the context of data assimilation. This ability has been exploited to define multiscale incremental data assimilation filtering schemes. This work has been recently published in Numerical Mathematics: Theory, Methods and Applications [14]. It is also described in Sébastien Beyou's PhD dissertation [11]. The development of an efficient GPU based version of this estimator recently started through the Inria ADT project FLUMILAB

6.1.2. 3D flows reconstruction from image data

Participants: Ioana Barbu, Kai Berger, Cédric Herzet, Etienne Mémin.

Our work focuses on the design of new tools for the estimation of 3D turbulent flow motion in the experimental setup of Tomo-PIV. This task includes both the study of physically-sound models on the observations and the fluid motion, and the design of low-complexity and accurate estimation algorithms. On the one hand, we investigate state-of-the-art methodologies such as "sparse representations" for the characterization of the observation and fluid motion models. On the other hand, we place the estimation problem into a probabilistic Bayesian framework and use state-of- the-art inference tools to effectively exploit the strong time-dependence on the fluid motion. In our previous work, we have focussed on the problem of reconstructing the particle positions from several two-dimensional images. Our approach was based on the exploitation of a particular family of sparse representation algorithms, leading to a good trade-off between performance and complexity. Moreover, we also tackled the problem of estimating the 3D velocity field of the fluid flow from two instances of reconstructed volumes of particles. Our approach was based on a generalization of the well-known Lucas-Kanade's motion estimator to 3D problems. A potential strength of the proposed approach is the possibility to consider a fully parallelized (and therefore very fast) hardware implementation. This year, we have focused on the design of new methodologies to jointly estimate the volume of particles and the velocity field from the received image data. Our approach is based on the minimization (with respect to both the position of the particles and the velocity field) of a cost function penalizing both the discrepancies with respect to a conservation equation and some prior estimates of particle positions. This work has led to one publication in an international conference (PIV13) [27] and one publication in a national conference (Fluvisu13) [31].

Since October 2013, with our new postdoctoral fellow Kai Berger, we have started a new direction of research targeting the volume reconstruction problem. In particular, we address the question of devising effective reconstruction procedures taking into account the limited computational budget available in practice. Our approach is based on the design of simple thresholding operators, allowing to reduce the dimension of the initial problem and amenable to fast parallel implementations.

6.1.3. Motion estimation techniques for turbulent fluid flows

Participants: Patrick Héas, Dominique Heitz, Cédric Herzet, Etienne Mémin.

In this study we have devised smoothing functional adapted to the multiscale structure of homogeneous turbulent flows. These regularization constraints ensue from a classical phenomenological description of turbulence. The smoothing is in practice achieved by imposing some scale-invariance principles between histograms of motion increments computed at different scales. Relying on a Bayesian formulation, an inference technique, based on likelihood maximization and marginalization of the motion variable, has been proposed to jointly estimate the fluid motion, the regularization parameters and a proper physical models. The performance of the proposed Bayesian estimator has been assessed on several image sequences depicting synthetic and real turbulent fluid flows. The results obtained in the context of fully developed turbulence show that an improvement in terms of small-scale motion estimation can be achieved as compared to classical motion estimator. This work, performed within a collaboration with Pablo Mininni from the university of Buenos Aires, have been published in the IEEE Transactions on Pattern Analysis And Machine Learning [22].

6.1.4. Wavelet basis for multiscale motion estimation

Participants: Patrick Héas, Cédric Herzet, Etienne Mémin.

In this study we focused on the implementation of a simple wavelet-based optical-flow motion estimator dedicated to the recovery of fluid motions. The wavelet representation of the unknown velocity field is considered. This scale-space representation, associated to a simple gradient-based optimization algorithm, sets up a natural multiscale/multigrid optimization framework for the optical flow estimation that can be combined to more traditional incremental multiresolution approaches. Moreover, a very simple closure mechanism, approximating locally the solution by high-order polynomials, is provided by truncating the wavelet basis at intermediate scales. This offers a very interesting alternative to traditional Particle Image Velocimetry techniques. As another alternative to this medium-scale estimator, we explored strategies to define estimation at finer scales. These strategies rely on the encoding of high-order smoothing functional on divergence free wavelet basis. This study has been published in the journal of Numerical Mathematics: Theory, Methods and Applications [19] and in the international Journal of Computer Vision [23]. This work has strongly benefited from a collaboration with Souleyman Kadri-Harouna (University of La Rochelle and who was formerly on a post-doctoral position in our team). The divergence free wavelets basis proposed in [24] constitutes the building blocks on which we have elaborated our wavelet based motion estimation solutions. We have otherwise pursue our collaboration with Chico university through the post-doc of Pierre Dérian on the GPU implementation of such motion estimator for Lidar data.

6.1.5. Sparse-representation algorithms

Participant: Cédric Herzet.

The paradigm of sparse representations is a rather new concept which turns out to be central in many domains of signal processing. In particular, in the field of fluid motion estimation, sparse representation appears to be potentially useful at several levels: i) it provides a relevant model for the characterization of the velocity field in some scenarios; ii) it plays a crucial role in the recovery of volumes of particles in the 3D Tomo-PIV problem.

Unfortunately, the standard sparse representation problem is known to be NP hard. Therefore, heuristic procedures have to be devised to try and access to the solution of this problem. Among the popular methods available in the literature, one can mention orthogonal matching pursuit, orthogonal least squares and the family of procedures based on the minimization of ℓ_p norms. In order to assess and improve the performance of these algorithms, theoretical works have been undertaken in order to understand under which conditions these procedures can succeed in recovering the "true" sparse vector.

This year, we have contributed to this research axis by deriving conditions of success for the algorithms mentioned above when some partial information is available about the position of the nonzero coefficients in the sparse vector. This paradigm is of interest in the Tomographic-PIV volume reconstruction problem: one can indeed expect volumes of particles at two successive instants to be pretty similar; any estimate of the position of the particles at one given instant can therefore serve as a prior estimate about their position at the next instant. The conditions of success of such procedure have been rigorously formalized in two publications in the IEEE Transactions on Information Theory [21], [26] and one publication in an international conference (SPARS13) [28].

6.2. Tracking, Data assimilation and model-data coupling

6.2.1. Stochastic filtering for fluid motion tracking

Participants: Sébastien Beyou, Anne Cuzol, Etienne Mémin.

Within the PhD thesis of Sébastien Beyou [11], we investigated the study of a recursive Bayesian filter for tracking velocity fields of fluid flows. We resort in this study to Monte-Carlo approximations based on the particle filtering paradigm. In particular, we investigated the use of the so-called ensemble Kalman filtering for fluid tracking problems. This kind of filters introduced for the analysis of geophysical fluids is based on the Kalman filter update equations. Nevertheless, unlike traditional Kalman filtering setting, the covariances of the estimation errors, required to compute the so-called Kalman gain, relies on an ensemble of forecasts. Such a process gives rise to a Monte-Carlo approximation for a family of non-linear stochastic filters enabling to handle state spaces of large dimension. The method we proposed can be seen as an extension of this technique that combines sequential importance sampling and the propagation law of an ensemble Kalman filter. This technique leads to an ensemble Kalman filter with an improved efficiency. Within this type of scheme, we have in particular investigated the introduction of a nonlinear direct image measurement operator. This modification of the filter provides very good results on 2D numerical and experimental flows even in the presence of strong noises. We assessed successfully its application to oceanic satellite images for the recovering of ocean streams. We have also studied the impact on the stochastic dynamics of auto-similar Gaussian noise mimicking statistical properties of turbulence and the introduction within an incremental ensemble analysis scheme of multiscale motion measurements. This work has been published in the Tellus A journal [15].

6.2.2. Stochastic filtering technique for the tracking of closed curves

Participant: Etienne Mémin.

We have studied a stochastic filtering technique for the tracking of closed curves along an image sequence. In that aim, we designed a continuous-time stochastic dynamics that allows us to infer inter-frame deformations. The curve is defined by an implicit level-set representation and the stochastic dynamics is expressed properly on the level-set function. It takes the form of a stochastic partial differential equation with a Brownian motion of low dimension. The evolution model we proposed combines local photometric information, deformations induced by the curve displacement and an uncertainty modeling of the dynamics. Specific choices of noise models and drift terms lead to an evolution law based on mean curvature as in classic level set methods, while other choices yield new evolution laws. The approach we propose is implemented through a particle filter, which includes color measurements characterizing the target and the background photometric probability densities respectively. The merit of this parameter free filter is demonstrated on various satellite image sequences depicting the evolution of complex geophysical flows. This work has been recently published in the Journal of Mathematical Imaging and Vision [13]. Let us note the method provides an empirical dynamical model learned recursively from a data flow. Its short time forecasting skills have been used in the context of weather-watch radar images within a fruitful collaboration with MeteoFrance.

6.2.3. Sequential smoothing for fluid motion

Participants: Anne Cuzol, Etienne Mémin.

In parallel to the construction of stochastic filtering techniques for fluid motions, we have proposed a new sequential smoothing method within a Monte-Carlo framework. This smoothing aims at reducing the temporal discontinuities induced by the sequential assimilation of discrete time data into continuous time dynamical models. The time step between observations can indeed be long in environmental applications for instance, and much longer than the time step used to discretize the model equations. While the filtering aims at estimating the state of the system at observations times in an optimal way, the objective of the smoothing is to improve the estimation of the hidden state between observation times. The method is based on a Monte-Carlo approximation of the filtering and smoothing distributions, and relies on a simulation technique of conditional diffusions. The proposed smoother can be applied to general non linear and multidimensional models. It has been applied to a turbulent flow in a high-dimensional context, in order to smooth the filtering results obtained from a particle filter with a proposal density built from an Ensemble Kalman procedure. This conditional simulation framework can also be used for filtering problem with low measurement noise. This has been explored through a collaboration with Jean-Louis Marchand (ENS Bretagne) in the context of vorticity tracking from image data.

6.2.4. *Stochastic fluid flow dynamics under uncertainty*

Participants: Etienne Mémin, Valentin Resseguier.

In this research axis we aim at devising Eulerian expressions for the description of fluid flow evolution laws under uncertainties. Such an uncertainty is modeled through the introduction of a random term that allows taking into account large-scale approximations or truncation effects performed within the dynamics analytical constitution steps. This includes for instance the modeling of unresolved scales interaction in large eddies simulation (LES) or in Reynolds average numerical simulation (RANS), but also uncertainties attached to non-uniform grid discretization. This model is mainly based on a stochastic version of the Reynolds transport theorem. Within this framework various simple expressions of the drift component can be exhibited for different models of the random field carrying the uncertainties we have on the flow. We aim at using such a formalization within image-based data assimilation framework and to derive appropriate stochastic versions of geophysical flow dynamical modeling. This formalization has been published in the journal Geophysical and Astrophysical Fluid Dynamics [25]. Numerical simulation on divergence free wavelets basis of 3D viscous Taylor-Green vortex and Crow instability have been performed within a collaboration with Souleymane Kadri-Harouna. First promising results have been published in the TSFP8 conference [30]. Besides, we explore in the context of Valentin Resseguier's PhD the extension of such framework to oceanic models and to satellite image data assimilation. This PhD thesis takes place within a fruitful collaboration with Bertrand Chapron (CERSAT/IFREMER).

6.2.5. *Free surface flows reconstruction and tracking*

Participants: Dominique Heitz, Etienne Mémin, Cordelia Robinson, Yin Yang.

Characterizing a free-surface flow (space and time-dependent velocity and geometry) given observations/measures at successive times is an ubiquitous problem in fluid mechanic and in hydrology. Observations can consist of e.g. measurements of velocity, or like in this work of measurements of the geometry of the free-surface. Indeed, recently developed depth/range sensors allow to capture directly a rough 3D geometry of surfaces with high space and time resolution. We have investigated the performance of the Kinect and have shown that it is likely to capture temporal sequences of depth observations of wave-like surfaces with wavelengths and amplitudes sufficiently small to characterize medium/large scale flows. Several data assimilation methods have been experimented and compared to estimate both time dependent geometry and displacement field associated to a free-surface flow from a temporal sequence of Kinect data. This study has been conducted on synthetic and real-world data. Finally, we explored the application of such techniques to hydrological applications. These results have been recently submitted to Journal of Computational Physics.

6.2.6. *Variationnal ensemble methods for data assimilation*

Participants: Dominique Heitz, Etienne Mémin, Cordelia Robinson, Yin Yang.

In this work, we aim at studying an ensemble based optimal control strategy for data assimilation. Such a formulation nicely combines the ingredients of ensemble Kalman filters and variational assimilation. In the same way as standard variational assimilation, it is formulated as the minimization of an objective function. However, similarly to ensemble filters, it introduces in its objective function an empirical ensemble-based background-error covariance and works in an off-line smoothing mode rather than sequentially like filtering approaches in a sequential filter. These techniques have the great advantage to avoid the introduction of tangent linear and adjoint models, which are necessary for standard incremental variational techniques. As the background error covariance matrix plays a key role in the variational process, our study particularly focuses on the generation of the analysis ensemble state with localization techniques. We compared the performances of both methods in different cases in which the system's component are fully observed or only partially. The comparisons have been leaded on the basis of a Shallow Water model.

6.2.7. Optimal control techniques for the coupling of large scale dynamical systems and image data

Participants: Dominique Heitz, Etienne Mémin, Cordelia Robinson.

This work aims at investigating the use of optimal control techniques for the coupling of Large Eddies Simulation (LES) techniques and 2D image data. The objective is to reconstruct a 3D flow from a set of simultaneous time resolved 2D image sequences visualizing the flow on a set of 2D plans enlightened with laser sheets. This approach will be experimented on shear layer flows and on wake flows generated on the wind tunnel of Irstea Rennes. Within this study we wish also to explore techniques to enrich large-scale dynamical models by the introduction of uncertainty terms or through the definition of subgrid models from the image data. This research theme is related to the issue of turbulence characterization from image sequences. Instead of predefined turbulence models, we aim here at tuning from the data the value of coefficients involved in traditional LES subgrid models or in longer-term goal to learn empirical subgrid models directly from image data. An accurate modeling of this term is essential for Large Eddies Simulation as it models all the non resolved motion scales and their interactions with the large scales.

We have pursued the first investigations on a 4DVar assimilation technique, integrating PIV data and Direct Numerical Simulation (DNS), to reconstruct two-dimensional turbulent flows. The problem we are dealing with consists in recovering a flow obeying Navier-Stokes equations, given some noisy and possibly incomplete PIV measurements of the flow. By modifying the initial and inflow conditions of the system, the proposed method reconstructs the flow on the basis of a DNS model and noisy measurements. The technique has been evaluated in the wake of a circular cylinder. It denoises the measurements and increases the spatiotemporal resolution of PIV time series. These results have been recently published in the Journal of Computational Physics [20]. A paper has been also recently published on the denoising aspect in the (PIV13) international conference [29]. Along the same line of studies the 3D case is ongoing. The goal consists here to reconstruct a 3D flow from a set of simultaneous time resolved 2D images of planar sections of the 3D volume. This work is mainly conducted within the PhD of Cordelia Robinson. The development of the variational assimilation code has been initiated within a collaboration with A. Gronskis, S. Laizé (lecturer, Imperial College, UK) and Eric Lamballais (institut P' Poitiers). A High Reynolds number simulation of the wake behind a cylinder has been recently performed within this collaboration.

6.2.8. Variational assimilation of images for large scale fluid flow dynamics with uncertainty

Participant: Etienne Mémin.

In this work we explore the assimilation of a large scale representation of the flow dynamics with image data provided at a finer resolution. The velocity field at large scales is described as a regular smooth components whereas the complement component is a highly oscillating random velocity field defined on the image grid but living at all the scales. Following this route we have started to assess the performance of a variational assimilation technique with direct image data observation. Preliminary encouraging results obtained for a wavelet-based 2D Navier Stokes implementation and images of a passive scalar transported by the flow have been obtained. Large-scale simulation under uncertainty for the 3D viscous Taylor-Green vortex flow have been carried out and show promising results of the approach.

6.2.9. Reduced-order models for flows representation from image data

Participants: Cédric Herzet, Etienne Mémin, Véronique Souchaud.

One of the possibilities to neglect the influence of some degrees of freedom over the main characteristics of a flow consists in representing it as a sum of K orthonormal spatial basis functions weighted with temporal coefficients. To determine the basis function of this expansion, one of the usual approaches relies on the Karhunen-Loeve decomposition (refered to as proper orthogonal decomposition – POD – in the fluid mechanics domain). In practice, the spatial basis functions, also called modes, are the eigenvectors of an empirical auto-correlation matrix which is built from “snapshots" of the considered physical process.

In this axis of work we focus on the case where one does not have a direct access to snapshots of the considered physical process. Instead, the POD has to be built from the partial and noisy observation of the physical phenomenon of interest. Instances of such scenarios include situations where real instantaneous vector-field snapshots are estimated from a sequence of images. We have been working on several approaches dealing with such a new paradigm. A first approach consists in extending standard penalized motion-estimation algorithms to the case where the sought velocity field is constrained to span a low-dimensional subspace. In particular, we have considered scenarios where the standard optical flow constraint (OFC) is no longer satisfied and one has therefore to resort to a Discrete Finite Difference (DFD) model. The non-linearity of the latter leads to several practical issues that we have addressed this year.

Within a collaboration with the University of Buenos Aires, we have also explored, a method that combines Proper Orthogonal Decomposition with a spectral technique to analyze and extract reduced order models of flows from time resolved data of velocity fields. This methodology, relying on the eigenfunctions of the Koopman operator, is specifically adapted to flows with quasi periodic orbits in the phase space. The technique is particularly suited to cases requiring a discretization with a high spatial and temporal resolution. The proposed analysis enables to decompose the flow dynamics into modes that oscillate at a single frequency. For each modes an energy content and a spatial structure can be put in correspondence. This approach has been assessed for a wake flow behind a cylinder at Reynolds number 3900 and has been recently published to the journal of Theoretical and Computational Fluid Dynamics [16]. The assessment of this method on oceanic model simulation data is on going.

6.3. Analysis and modeling of turbulent flows

6.3.1. Hot-wire anemometry at low velocities

Participant: Dominique Heitz.

A new dynamical calibration technique has been developed for hot-wire probes. The technique permits, in a short time range, the combined calibration of velocity, temperature and direction calibration of single and multiple hot-wire probes. The calibration and measurements uncertainties were modeled, simulated and controlled, in order to reduce their estimated values. Based on a market study the french patent application has been extended this year to a Patent Cooperation Treaty (PCT) application.

6.3.2. Numerical and experimental image and flow database

Participant: Dominique Heitz.

The goal was to design a database for the evaluation of the different techniques developed in the Fluminance group. The main challenge was to enlarge a database mainly based on two-dimensional flows, with three-dimensional turbulent flows. New synthetic image sequences based on homogeneous isotropic turbulence and on circular cylinder wake have been provided. These images have been completed with real image sequences based on wake and mixing layers flows. This new database provides different realistic conditions to analyse the performance of the methods: time steps between images, level of noise, Reynolds number, large-scale images. A Wake flow at high Reynolds number has been also simulated on one of the IDRISS super computer. This simulation, whose results analysis is on going, has been performed within a collaboration with Sylvain Laizet (Imperial College).

6.4. Visual servoing approach for fluid flow control

6.4.1. Minimization of the kinetic energy density in the 2D plane Poiseuille flow

Participants: Christophe Collewet, Xuan Quy Dao.

This works concerns the PhD thesis of Xuan-Quy Dao. This year we have focused on a way to ensure a strict decreasing of the kinetic energy density. In that purpose, we have first proposed an approach to increase the controlled degrees of freedom. Indeed, the classical way to model this flow leads to only two degrees of freedom. With so few degrees of freedom it is obviously impossible to reach high desired performances as the strict minimization of the kinetic energy density. This way to proceed leads to a better minimization of the kinetic energy density. We have also proposed on approach based on a local decoupling of the controlled degree of freedom of the system so that an exponential decoupled decrease of each components of the state vector is locally obtained. This work has been presented at the CFM conference (Congrés Français de Mécanique) [32].

6.4.2. Control behind a backward-facing step

Participant: Christophe Collewet.

This work is performed in the context of the PhD thesis of Nicolas Gautier from ESCPI in collaboration with J.L. Aider. The separated flow downstream a backward-facing step is studied using visual information for feedback. More precisely, flow velocity fields are computed from a real time optical flow algorithm. The control law we used is a simple PID controller. Even a better control law could be used, this study validates that visual servo control is an effective approach to control a flow.

6.4.3. Control of systems described by partial differential equations

Participants: Tudor-Bogdan Airimitoaie, Christophe Collewet.

This work concerns principally the post-doctoral research of Tudor-Bogdan Airimiţoaie. It aims at controlling continuously evolving systems described by partial differential equations (PDEs). This is relevant in the context of the Fluminance team because fluid flows are infinite dimensional systems and can be rigorously described only through PDEs. In spite of this, practical approaches of flow control are based on low order numerical implementation relying on space and time discretization of the continuous system. This implies to setup strategies for model reduction that must be then in return properly understood with respect to the convergence of the control law. For finite dimensional implementations, one of the research directions pursued concerns the study on the benefit of increasing the controlled degrees of freedom (see the work of Xuan-Quy Dao). Another research direction, started recently, consists in improving control by using real-time estimation of a finite number of parameters related to the original infinite dimensional system. Indeed, this opens the possibility of improving performances by using more advanced robust linear parametric varying (LPV) control techniques existing in the literature. Two conference papers on these works have been submitted at the "7th AIAA Flow Control Conference".

GENSCALE Project-Team

6. New Results

6.1. NGS methodology

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Anaïs Gouin, Fabrice Legeai.

- **Efficient Kmer counting**: Counting all the substrings of length k (k-mers) in DNA/RNA sequencing reads is the preliminary step of many bioinformatics applications. However, state of the art k-mer counting methods require that a large data structure resides in memory. Such structure typically grows with the number of distinct k-mers to count. We have developed a new streaming algorithm for that purpose which only requires a fixed user-defined amount of memory and disk space. This approach realizes a memory, time and disk trade-off. DSK is the first approach that is able to count all the 27-mers of a human genome dataset using only 4.0 GB of memory and moderate disk space (160 GB), in 17.9 h. DSK can replace a popular k-mer counting software (Jellyfish) on small-memory servers. [24]
- **Questionning the classical re-sequencing analyses approach**: Classical re-sequencing analyses are based on a first step of read mapping, then only mapped reads are taken into account in following analyses such as variant calling. We investigated the sources of unmapped reads in aphid re-sequencing data of 33 individuals, and we demonstrated that these reads contain valuable information that should not be discarded as usually done in such analyses. We proposed also an approach to extract this information, based on assembly and re-mapping. [34]
- **Repeat detection** A new algorithm was developed for detecting long similar fragments occurring at least twice in a set of biological sequences. The problem becomes computationally challenging when the frequency of a repeat is allowed to increase and when a non-negligible number of insertions, deletions and substitutions are allowed. The proposed algorithm, called Rime (for Repeat Identification: long, Multiple, and with Edits) performs this task, and manages instances whose size and combination of parameters cannot be handled by other currently existing methods. To the best of our knowledge, Rime is the first algorithm that can accurately deal with very long repeats (up to a few thousands), occurring possibly several times, and with a rate of differences (substitutions and indels) allowed among copies of a same repeat of 10-15% or even more. [17]

6.2. NGS applications

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Fabrice Legeai.

- **Participation to an international competition of assembly**: The process of generating raw genome sequence data continues to become cheaper, faster, and more accurate. However, assembly of such data into high-quality finished genome sequences remains challenging. Many genome assembly tools are available, but they differ greatly in terms of their performance and in their final output. More importantly, it remains largely unclear how to best assess the quality of assembled genome sequences. In this context, we have participated to the Assemblathon-2 competitions, which purpose was to assess current state-of-the-art methods in genome assembly. Globally, the cumulative z-scores of different assembly criteria set our assembly strategy in the 4th position compared to other competitors (21 groups). [12]
- **Assembly on Raspberri Pi**: Current Assembly tools require computers with large memory configuration. In order to demonstrate the efficiency of our low memory footprint assembly tools, we assemble the genome of C. Elegans (100 Mbp) on the raspberry PI computer, a small system equipped with only 512 MB RAM and 32 GB flash drive. [42]

- **SNP detection on the tick** We took part of a population genetic study on the tick species *Ixodes ricinus*, the main vector species of human and animal vector-borne diseases in Europe. In this framework, we proposed the first identification of a set of SNPs isolated from the genome of *I. ricinus*, by applying, among others a new tool developed in the GenScale team: discoSnp. The main advantage of this tool is to be able to detect SNPs without the use of a reference genome, which is crucially lacking for the tick species. Among the detected SNPs, 384 were selected, according to their minimal and maximal coverage and context sequences for experimental validation. Among them, 368 (95.8%) were biologically validated, demonstrating the precision of discoSNP.[23]
- **NGS analyses on insect models** We achieved the transcriptome assembly and analyzed the differential expression of an important noctuid pest. [22], [18]. Using gene expression data (RNA-Seq) in males, sexual females and asexual females of the pea aphid, we confirm theoretical models suggesting that the evolution of sex-biased gene expression may restrict the product of a sexually antagonistic allele to the sex it benefits.[19]
- **Genome sequencing and annotation**: We participated in the sequencing and annotation of several bacterial species of the Mollicute group. These bacteria are important pathogens of ruminants. The sequencing and annotation of their genomes confirmed their pathogenic features and phylogenetic location in the tree of Mollicutes. This is the first step before comparative genome analyses to unravel the genetic basis of mycoplasma pathogenicity and host specificity. [15], [16], [21]

6.3. HPC and parallelism

Participants: Dominique Lavenier, Rumen Andonov, Guillaume Chapuis, François Moreews, Charles Deltel.

- **Improving time performances of Mapping quantitative trait loci (QTL)** : we have developed a fast implementation of QTLMap, which takes advantage of the data parallel nature of the problem by offsetting heavy computations to a graphics processing unit (GPU). This new implementation performs up to 75 times faster than the previous multicore implementation, while maintaining the same results and level of precision . This speedup allows one to perform more complex analyses, such as linkage disequilibrium linkage analyses (LDLA) and multiQTL analyses, in a reasonable time frame. [13]
- **Integration of parallelism in bioinformatics workflows**: We propose a Model-Driven Architecture approach for capturing the complete design process of bioinformatics workflows. This approach is applied to graphical workflow editors and allows to quickly convert a workflow prototype in a parallel implementation. This work can have an impact on the way bioinformaticians implement their analysis and increase their productivity.[30]
- **Parallel assembly on FPGAs**: This research work proposes a method to reduce the overall time for assembly by using pre-processing of the short read data on FPGAs and processing its output using Velvet. We demonstrate significant speed-ups with slight or no compromise on the quality of the assembled output.[32]
- **All-Pairs Shortest Paths with multi-GPU** We propose a new algorithm for the All-Pairs Shortest Paths problem for graphs with good partitioning properties and its multi-GPU implementation. Our implementation targets large graphs (up to 10^6 vertices) and allows graphs with negative edges to be computed. [35]

6.4. Protein structures

Participants: Rumen Andonov, Guillaume Chapuis, Dominique Lavenier, Mathilde Le Boudic-Jamin, Antonio Mucherino, Douglas Goncalves.

- **A book on distance geometry problems (DGP).** This is a collection of invited papers on the topic "distance geometry" [38]. Among the other contributions, it contains a survey on "distance geometry" and "structural biology", which tries to function as a bridge between two scientific communities: computer science and biology. It presents some recent developments in the field by

using a language common to the two communities [37]. In another contribution, the complexity of the DGP is discussed: even if this problem is NP-hard in general, we noticed a polynomial complexity on instances of DGP related to protein conformations (in the case all the available distances are exact)[36].

- **DGP with interval data.** In our preliminary works on the discretization of the Distance Geometry Problem (DGP), we considered instances where all distances were supposed to be exactly known. When biological molecules are concerned, however, this is not generally the case. We worked therefore for considering the full-atom representation of the protein backbone, where some of the distances are subject to uncertainty within a given nonnegative interval. We showed that the discretization is still possible in this case, and proposed the iBP algorithm to solve the discretized DGP. [20]
- **New pruning device for DGP.** After the discretization, DGPs can be solved by a branch-and-prune (BP) algorithm, which is potentially able to enumerate the entire solution set. This solution set, however, can be very large for some instances, while only the most energetically stable conformations are of interest. We worked therefore for integrating the BP algorithm with two new energy-based pruning devices. Our computational experiments showed that the newly added pruning devices were actually able to improve the performance of the algorithm, as well as the quality (in terms of energy) of the conformations in the solution set. [28]
- **Discretization orders for the DGP.** The main assumption that allows for the discretization of DGPs is strongly based on the order in which the atoms of the molecule are considered. The "natural" order of the atoms in the amino acid chain does not always allow for the discretization. We tried to find discretization orders in several ways, based on different approaches. In [31], we extended a previously proposed greedy algorithm that is able to deal with interval data (inexact distances). In [27], we handcrafted some discretization orders for the side chains of the amino acids involved in the protein synthesis. In [29], we proposed a heuristic, which outperforms, on large instances, the greedy algorithm previously proposed.
- **DGP with Clifford Algebra** The BP algorithm for the DGP is based on a search on the tree, where nodes of the tree belonging to a common layer provide the possible positions for the same atom of the molecule. When interval data are given, a curve in 3d (containing the possible positions for the atom) can be associated to one of such nodes. Since it is generally not necessary to have protein conformations with a precision higher than 1A, sample points on these curves can be chosen. The way to choose these sample points is not, however, a simple task. This is the reason why we are trying to make this selection process adaptive, by exploiting Clifford Algebra to this purpose. Preliminary studies in this direction were presented in [25]
- **Parallel seed-based approach to protein structure similarity detection** We have developed a new parallel heuristic-based approach to structural similarity detection between proteins that discovers multiple pairs of similar regions. We prove that returned alignments have RMSDc and RMSDd lower than a given threshold. Computational complexity is addressed by taking advantage of both fine- and coarse-grain parallelism. [26]
- **Datamining.** The selection of features that describe samples in sets of data is a typical problem in data mining. A crucial issue is to select a maximal set of pertinent features, because the scarce knowledge of the problem under study often leads to consider features which do not provide a good description of the corresponding samples. The concept of consistent biclustering of a set of data has been introduced to identify such a maximal set. The problem can be modeled as a 0–1 linear fractional program, which is NP-hard. We reformulated this optimization problem as a bilevel program, and we proposed a heuristic for its solution [39].

HYBRID Project-Team

6. New Results

6.1. 3D interactive techniques

6.1.1. Navigating in virtual environments with omnidirectional rendering

Participants: Jérôme Ardouin [contact], Anatole Lécuyer [contact], Maud Marchal.

The "FlyVIZ" enables humans to experience a real-time 360o vision of their surroundings for the first time. The visualization device combines a panoramic image acquisition system (positioned on top of the user's head) with a Head-Mounted Display (HMD). The omnidirectional images are transformed to fit the characteristics of HMD screens. As a result, the user can see his/her surroundings, in real-time, with 360o images mapped into the HMD field-of- view.

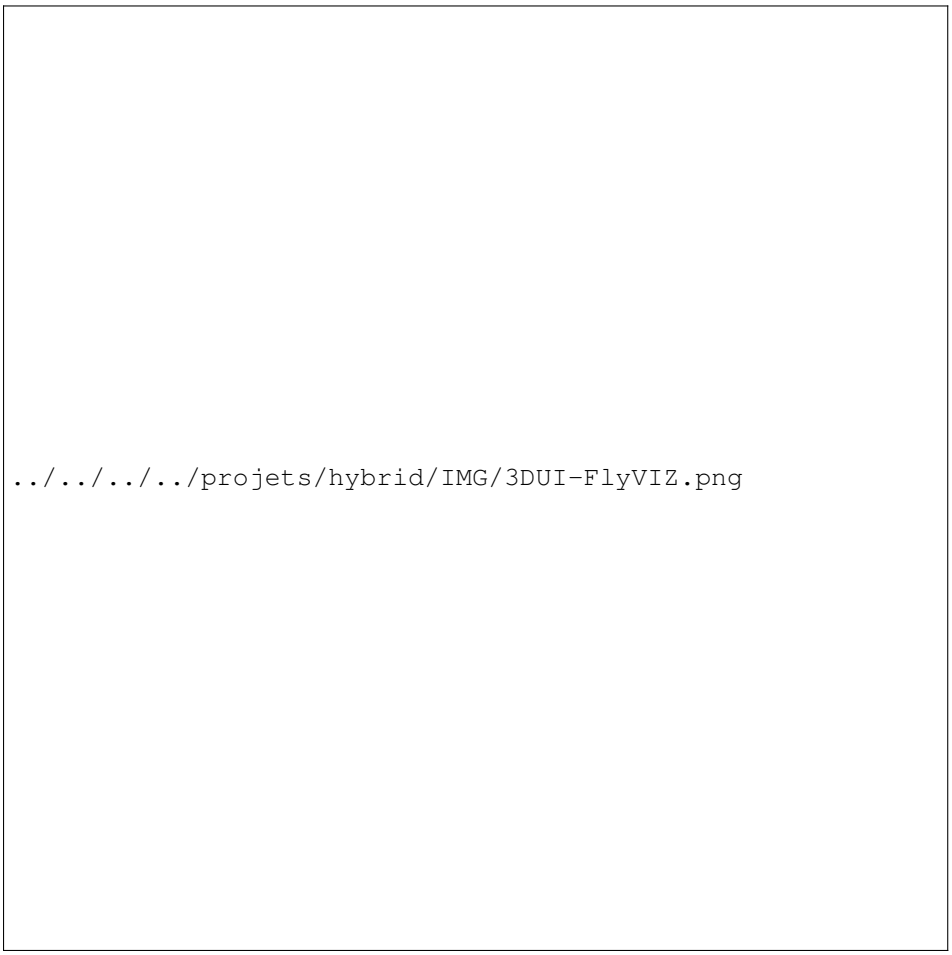


Figure 2. The "FlyVIZ" enables humans to experience in real-time a 360-degree vision of their surroundings.

In order to safely simulate and evaluate our approach, we designed and evaluated [27] several visualization techniques, for navigating in virtual environments (VE). We have conducted an evaluation of different methods compared to a rendering method of reference, i.e. a perspective projection, in a basic navigation task. Our results confirm that using any omnidirectional rendering method could lead to more efficient navigation in terms of average task completion time. Among the different 360o projection methods, the subjective preference was significantly given to a cylindrical projection method (equirectangular). Taken together, our results suggest that omnidirectional rendering could be used in virtual reality applications in which fast navigation or full and rapid visual exploration are important. They pave the way to novel kinds of visual cues and visual rendering methods in virtual reality. This work was a collaboration with the Lagadic team (Inria Rennes).

6.1.2. Advances in locomotion interfaces for virtual environments

Participants: Anatole Lécuyer [contact], Maud Marchal [contact], Bruno Arnaldi.

Navigation, a fundamental task in Virtual Reality (VR), is greatly influenced by the locomotion interface being used, by the specificities of input and output devices, and by the way the virtual environment is represented. No matter how virtual walking is controlled, the generation of realistic virtual trajectories is absolutely required for some applications, especially those dedicated to the study of walking behaviors in VR, navigation through virtual places for architecture, rehabilitation and training.

First, we have studied the realism of unconstrained trajectories produced during virtual walking.We proposed a comprehensive evaluation framework consisting on a set of trajecto-graphical criteria and a locomotion model to generate reference trajectories [16]. We considered a simple locomotion task where users walk between two oriented points in space. The travel path was analyzed both geometrically and temporally in comparison to simulated reference trajectories. This work was a collaboration with the Mimetic team (Inria Rennes).

Secondly, we have introduced novel "Camera Motions" (CMs) to improve the sensations related to locomotion in virtual environments (VE) [26]. Traditional CMs are artificial oscillating motions applied to the subjective viewpoint when walking in the VE, and they are meant to evoke and reproduce the visual flow generated during a human walk. Our novel CMs are: (1) multistate, (2) personified, and (3) they can take into account the topography of the virtual terrain. In addition, they can then take into account avatar's fatigue and recuperation, and the topography for updating visual CMs accordingly. Taken together, our results suggest that our new CMs could be introduced in Desktop VR applications involving first-person navigation, in order to enhance sensations of walking, running, and sprinting, with potentially different avatars and over uneven terrains, such as for training, virtual visits or video games.

6.1.3. 3D manipulation of virtual objects: 3-Point++

Participants: Thierry Duval [contact], Thi Thuong Huyen Nguyen.

Manipulation in immersive Virtual Environments (VEs) is often difficult and inaccurate because humans have difficulty in performing precise positioning tasks or in keeping the hand motionless in a particular position without any help of external devices or haptic feedback. To address this problem, we proposed a set of four manipulation points attached to objects (called a 3-Point++ tool, including three handle points and their barycenter), by which users can control and adjust the position of objects precisely [39]. By determining the relative position between the 3-Point++ tool and the objects, and by defining different states of each manipulation point (called locked/unlocked or inactive/active), these points can be freely configured to be adaptable and flexible to enable users to manipulate objects of varying sizes in many kinds of positioning scenarios.

6.1.4. A survey of 3D object selection techniques for virtual environments

Participant: Ferran Argelaguet Sanz [contact].

Computer graphics applications controlled through natural gestures are gaining increasing popularity these days due to recent developments in low-cost tracking systems and gesture recognition technologies. Although interaction techniques through naturalgestures have already demonstrated their benefits in manipulation, navigation and avatar-control tasks, effective selection with pointing gestures remains an open problem. We surveyed the state-of-the-art in 3D object selection techniques [13]. We reviewed important findings in human control models, analyze major factors influencing selection performance, and classify existing techniques according to a number of criteria. Unlike other components of the application's user interface, pointing techniques need a close coupling with the rendering pipeline, introducing new elements to be drawn, and potentially modifying the object layout and the way the scene is rendered. Conversely, selection performance is affected by rendering issues such as visual feedback, depth perception, and occlusion management. We thus reviewed existing literature paying special attention to those aspects in the boundary between computer graphics and human computer interaction.

6.1.5. Novel pseudo-haptic based interfaces

Participants: Pierre Gaucher, Ferran Argelaguet Sanz, Anatole Lécuyer [contact], Maud Marchal.

Pseudo-haptics is a technique meant to simulate haptic sensations using visual feedback and properties of human visuo-haptic perception. In this course of action, we have extended its usage for gestural interfaces [32] and exploring its usage for the simulation of the local elasticity of images [].

Interacting with virtual objects through free-hand gestures do not allow users to perceive the physical properties of virtual objects. To provide enhanced interaction, we explored how the usage of a pseudo-haptic approach could be introduced while interacting with a 3D Carrousel [32]. In our approach, which is envisioned for showcasting purposes, virtual products are presented using a 3D carousel augmented with physical behavior and a pseudo-haptic effect aiming to attract the user to specific items. The user, through simple gestures, controls the rotation of the carousel, and can select, examine and manipulate the objects presented. Several demos can be tested on-line at Hybrid website.

Secondly, we have introduced the Elastic Images, a novel pseudo-haptic feedback technique which enables the perception of the local elasticity of images without the need of any haptic device []. The proposed approach focuses on whether visual feedback is able to induce a sensation of stiffness when the user interacts with an image using a standard mouse. The user, when clicking on a Elastic Image, is able to deform it locally according to its elastic properties. A psychophysical experiment was conducted to quantify this novel pseudo-haptic perception and determine its perceptual threshold (or its Just Noticeable Difference). The results showed that users were able to recognize up to eight different stiffness values with our method and confirmed that it provides a perceivable and exploitable sensation of elasticity.

6.1.6. Experiencing the past in virtual reality

Participant: Valérie Gouranton [contact].

We designed a public experience and exhibition organized during the French National Days of Archaeology. This was the result of an interdisciplinary collaboration between archaeologists and computer scientists, centered on the immersive virtual reality platform Immersia, a node of the European Visionair project. This public exhibition had three main goals: (i) presenting our interdisciplinary collaboration, (ii) communicating on the scientific results of this collaboration, and (iii) offering an immersive experience in the past for visitors. In [33] we could present the scientific context of the event, its organization, and a discussion on feedbacks.

In the frame of the CNPAO project (section 8.1.3) we have also worked on the reconstitution of six archaeological sites located in the west of France ranging from prehistory to the Middle Ages: the Cairn of Carn Island, the covered pathway of Roh Coh Coet, the GohMin Ru megalithic site, the gallo-roman mansion of Vanesia, the keep of the Château de Sainte-Suzanne, the Porte des Champs of the Château d'Angers. Other proposals are currently under study [29].

6.1.7. Perception of affordances in virtual reality

Participants: Anatole Lécuyer [contact], Maud Marchal.

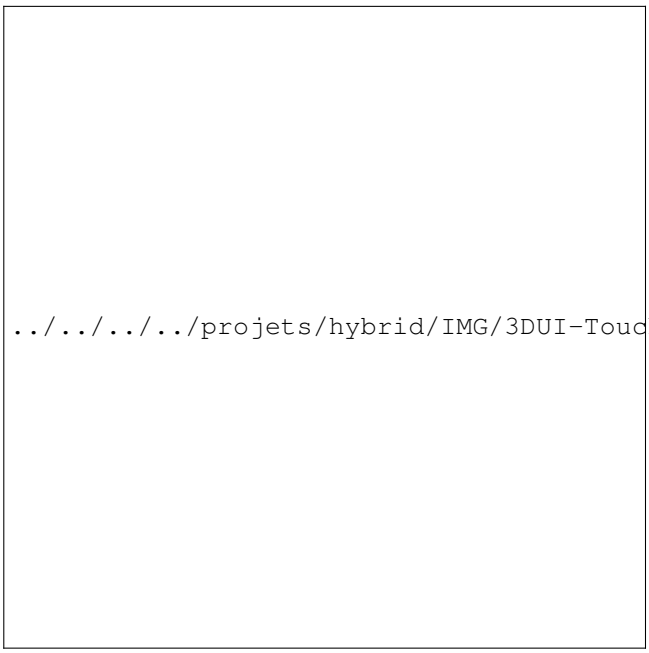


Figure 3. "Touching the past" experience during the French National Days of Archaeology.

The perception of affordances could be a potential tool for sensorimotor assessment of physical presence, that is, the feeling of being physically located in a virtual place. We have evaluated the perception of affordances for standing on a virtual slanted surface [25]. Participants were asked to judge whether a virtual slanted surface supported up right stance. The objective was to evaluate if this perception was possible in virtual reality (VR) and comparable to previous works conducted in real environments. We found that the perception of affordances for standing on a slanted surface in virtual reality is possible and comparable (with an underestimation) to previous studies conducted in real environments. We also found that participants were able to extract and to use virtual information about friction in order to judge whether a slanted surface supported an upright stance. Finally, results revealed that the person's position on the slanted surface is involved in the perception of affordances for standing on virtual grounds. Taken together, our results show quantitatively that the perception of affordances can be effective in virtual environments, and influenced by both environmental and person properties. Such a perceptual evaluation of affordances in VR could guide VE designers to improve their designs and to better understand the effect of these designs on VE users.

6.2. Haptic Feedback and Physical Simulation

6.2.1. Haptic feedback to improve audiovisual experience

Participants: Fabien Danieau, Anatole Lécuyer [contact].

Haptics have been employed in a wide set of applications ranging from teleoperation and medical simulation to arts and design, including entertainment, aircraft simulation and virtual reality. As for today, there is also a growing attention from the research community on how haptic feedback can be integrated with profit to audiovisual systems. We have first reviewed [18] the techniques, formalisms and key results on the enhancement of audiovisual experience with haptic feedback. We first reviewed the three main stages in the pipeline which are (i) production of haptic effects, (ii) distribution of haptic effects and (iii) rendering of haptic

effects. We then highlighted the strong necessity for evaluation techniques in this context and discuss the key challenges in the field. By building on technology and results from virtual reality, and tackling the specific challenges in the enhancement of audiovisual experience with haptics, we believe the field presents exciting research perspectives for which financial and societal stakes are significant.

We have then developped a novel approach called HapSeat for simulating motion sensations in a consumer environment.Multiple force-feedbacks are applied to the seated user's body to generate a 6DoF sensation of motion while experiencing passive navigation as illustrated Figure 4 . A set of force-feedback devices such as mobile armrests or headrests are arranged around a seat so that they can apply forces to the user. The forces are computed consistently with the visual content (visual acceleration) in order to generate motion sensations. This novel display device has been patented and was demonstrated this year at ACM SIGGRAPH 2013 Emerging-Technologies [55], and ACM CHI 2013 Interactivity [54].

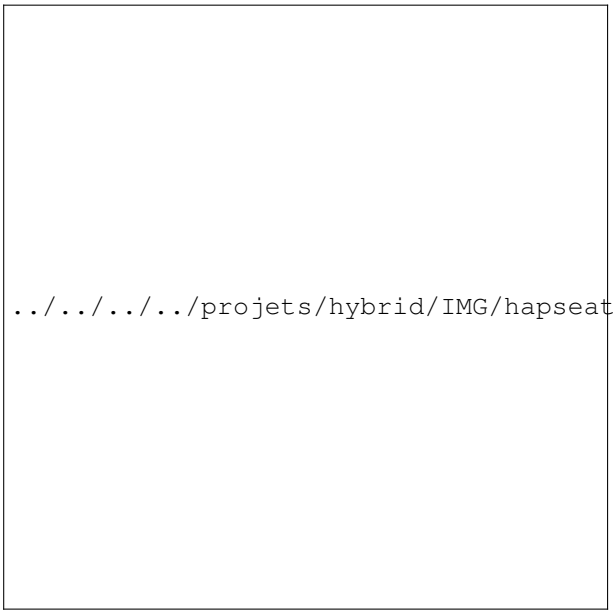


Figure 4. The HapSeat device: force-feedback is applied on the user's body with mobile armrests or headrests in order to generate motion sensations that are consistent with the visual content.

This work was a collaboration with the Mimetic team (Inria Rennes).

6.2.2. Vibrotactile rendering of splashing fluids

Participants: Anatole Lécuyer, Maud Marchal [contact].

Compelling virtual reality scenarios involving physically based virtual materials have been demonstrated using hand- based and foot-based interaction with visual and vibrotactile feedback. However, some materials, such as water and other fluids, have been largely ignored in this context. For VR simulations of real-world environments, the inability to include interaction with fluids is a significant limitation. Potential applications include improved training involving fluids, such as medical and phobia simulators, and enhanced user experience in entertainment, such as when interacting with water in immersive virtual worlds. We introduced the use of vibrotactile feedback as a rendering modality for solid-fluid interaction, based on the physical processes that generate sound during such interactions [15]. This rendering approach enables the perception

of vibrotactile feedback from virtual scenarios that resemble the experience of stepping into a water puddle or plunging a hand into a volume of fluid.

6.2.3. Six-DoF haptic interaction with fluids, solids, and their transitions

Participants: Anatole Lécuyer, Maud Marchal [contact].

Haptic interaction with different types of materials in the same scene is a challenging task, mainly due to the specific coupling mechanisms that are usually required for either fluid, deformable or rigid media. Dynamically-changing materials, such as melting or freezing objects, present additional challenges by adding another layer of complexity in the interaction between the scene and the haptic proxy. We have addressed these issues through a common simulation framework, based on Smoothed-Particle Hydrodynamics, and enable haptic interaction simultaneously with fluid, elastic and rigid bodies, as well as their melting or freezing [30]. We introduced a mechanism to deal with state changes, allowing the perception of haptic feedback during the process, and a set of dynamic mechanisms to enrich the interaction through the proxy. We decouple the haptic and visual loops through a dual GPU implementation. An initial evaluation of the approach was performed through performance and feedback measurements, as well as a small user study assessing the capability of users to recognize the different states of matter they interact with.

6.2.4. Bimanual haptic manipulation

Participants: Anatole Lécuyer [contact], Maud Marchal [contact], Anthony Talvas.

Bimanual haptics is a specific kind of multi-finger interaction that focuses on the use of both hands simultaneously. Several haptic devices enable bimanual haptic interaction, but they are subject to a certain number of limitations for interacting with virtual environments (VEs), such as workspace size issues or manipulation difficulties, notably with single-point interfaces. Interaction techniques exist to overcome these limitations and allow users to perform specific two-handed tasks, such as the bimanual exploration of large VEs and grasping of virtual objects. We have proposed an overview of the current limitations in bimanual haptics and the interaction techniques developed to overcome them. Novel techniques based on the Bubble technique are more specifically presented, with a user evaluation that assesses their efficiency. These include bimanual workspace extension techniques as well as techniques to improve the grasping of virtual objects with dual single-point interfaces. This work was published as a chapter in a book on “Multi-finger Haptic Interaction” [51].

6.2.5. The god-finger method

Participants: Anatole Lécuyer, Maud Marchal [contact], Anthony Talvas.

In physically-based virtual environments, interaction with objects generally happens through contact points that barely represent the area of contact between the user’s hand and the virtual object. This representation of contacts contrasts with real life situations where our finger pads have the ability to deform slightly to match the shape of a touched object. We have proposed a method called god-finger to simulate a contact area from a single contact point determined by collision detection, and usable in a rigid body physics engine [42]. The method uses the geometry of the object and the force applied to it to determine additional contact points that will emulate the presence of a contact area between the user’s proxy and the virtual object. It could improve the manipulation of objects by constraining the rotation of touched objects in a similar manner to actual finger pads. An implementation in a physics engine shows that the method could make for more realistic behaviour when manipulating objects while keeping high simulation rates. This work was presented at IEEE 3DUI Symposium 2013 and has received the best technote award [42].

6.2.6. Collision detection for fracturing rigid bodies

Participant: Maud Marchal [contact].

Figure 5. 3D interaction techniques for bimanual haptic manipulation.

In complex scenes with many objects, collision detection plays a key role in the simulation performance. This is particularly true for fracture simulation, where multiple new objects are dynamically created. We have proposed novel algorithms and data structures for collision detection in real-time brittle fracture simulations [21]. We build on a combination of well-known efficient data structures, namely distance fields and sphere trees, making our algorithm easy to integrate on existing simulation engines. We proposed novel methods to construct these data structures, such that they can be efficiently updated upon fracture events and integrated in a simple yet effective self-adapting contact selection algorithm. Altogether, we drastically reduced the cost of both collision detection and collision response. We have evaluated our global solution for collision detection on challenging scenarios, achieving high frame rates suited for hard real-time applications such as video games or haptics. Our solution opens promising perspectives for complex brittle fracture simulations involving many dynamically created objects.

Figure 6. Example of brittle fracture with collision detection.

This work was a collaboration with the Mimetic team (Inria Rennes).

6.2.7. Collision detection with high performance computing on GPU

Participants: Bruno Arnaldi, Valérie Gouranton [contact], François Lehericey.

We have first proposed IRTCD, a novel Iterative Ray-Traced Collision Detection algorithm that exploits spatial and temporal coherency. Our approach uses any existing standard ray-tracing algorithm and we propose an iterative algorithm that updates the previous time step results at a lower cost with some approximations. Applied for rigid bodies, our iterative algorithm accelerate the collision detection by a speedup up to 33 times compared to non-iterative algorithms on GPU [34].

Then, we have presented two methods to efficiently control and reduce the interpenetration without noticeable computation overhead. The first method predicts the next potentially colliding vertices. These predictions are used to make our IRTCD algorithm more robust to the approximations, therefore reducing the errors up to 91%. We also present a ray re-projection algorithm that improves the physical response of ray-traced collision detection algorithm. This algorithm also reduces, up to 52%, the interpenetration between objects in a virtual environment. Our last contribution showed that our algorithm, when implemented on multi-GPUs architectures, is far faster [35].

Finally, we proposed a distributed and anticipative model for collision detection and propose a lead for distributed collision handling, two key components of physically-based simulations of virtual environments. This model is designed to improve the scalability of interactive deterministic simulations on distributed systems such as PC clusters. Our main contribution consists of loosening synchronism constraints in the collision detection and response pipeline to allow the simulation to run in a decentralized, distributed fashion.

Figure 7. Real-time simulation of iterative ray-traced collision detection algorithm.

We could show the potential for distributed load balancing strategies based on the exchange of grid cells, and explain how anticipative computing may, in cases of short computational peaks, improve user experience by avoiding frame-rate drop-downs [31].

6.3. Brain-Computer Interfaces and Virtual Environments

6.3.1. Multi-user BCI video game

Participant: Anatole Lécuyer [contact].

How can we connect two brains to a video game by means of a BCI, and what will happen when we do so? How will the two users behave, and how will they perceive this novel common experience? We have created a multi-user videogame called "BrainArena" in which two users can play a simple football game by means of two BCIs [14], as illustrated Figure 8 . They can score goals on the left or right side of the screen by simply imagining left or right hand movements. To add another interesting element , the gamers can play in a collaborative manner (their two mental activities are combined to score in the same goal), or in a competitive manner (the gamers must push the ball in opposite directions). Two experiments were conducted to evaluate the performance and subjective experience of users in the different conditions. Taken together our results suggest that multi-user BCI applications can be operational, effective, and more engaging for participants.

This work was a collaboration with the Potioc team (Inria Bordeaux).

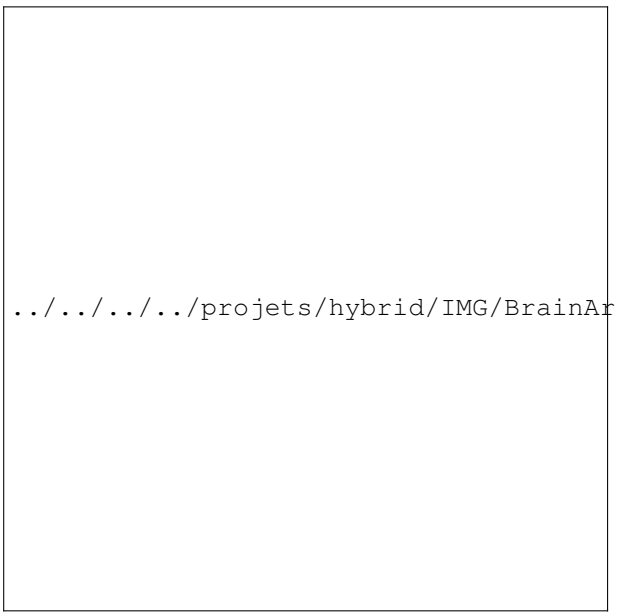


Figure 8. Multi-user football videogame in which two players can score goals to the left or right by imagining left or right hand movements. The users can play together using their brain activities either in a collaboration mode (same goal) or in a competitive mode (one versus the other).

6.3.2. Contextual SSVEP-based BCI control

Participants: Jozef Legény, Anatole Lécuyer [contact].

One main disadvantage of Brain-Computer Interfaces is that they are not completely reliable. In order to increase BCI performances, some adjustments can be made on low levels, such as signal processing and on high levels by modifying the controller paradigm. We have explored a novel, context-dependent, approach for SSVEP-based BCI controller [22]. This controller uses two kinds of behaviour alternation, commands can be added and removed if their use is irrelevant to the context or the actions resulting from their activation can be weighted depending on the likeliness of the actual intention of the user. This controller has been integrated within a BCI computer game and its influence in performance and mental workload has been addressed through a pilot experiment. Preliminary results have shown a workload reduction and performance improvement with the context-dependent controller while keeping the engagement levels untouched.

This work was a collaboration with the Universidad de Jaen (Spain).

6.3.3. Can we use a BCI and manipulate a mouse at the same time?

Participants: Jonathan Mercier-Ganady, Anatole Lécuyer [contact], Maud Marchal.

In most setups using a BCI, the user is explicitly asked to remain as motionless as possible, since muscular activity is commonly admitted to add noise and artifacts in brain electrical signals. Thus, as for today, people have been rarely let using other classical input devices such as mice or joysticks simultaneously to a BCI-based interaction. We have conducted an experimental study on the influence of manipulating an input device such as a standard computer mouse on the performance of a BCI system [37]. The study uses a simple virtual environment inspired by the well-known Pac-Man videogame and based on BCI and mouse controls. As expected the BCI performance was found to slightly decrease in presence of motor activity. However, we found that the BCI could still be successfully used in all conditions, even in presence of a highly-demanding mouse manipulation. These promising results pave the way to future experimental studies with more complex mental and motor activities, but also to novel 3D interaction paradigms that could mix BCI and other input devices for virtual reality and videogame applications.

6.3.4. Adaptive VR simulators combining visual, haptic, and BCIs

Participants: Anatole Lécuyer [contact], Maud Marchal.

What if the next generation of virtual reality simulators would take into account a novel user's input: his/her mental state, as measured with electrodes and Brain-Computer Interfaces ? This would lead to adaptive simulators that could match the "hidden" expectations of the user optimally? We have initiated and illustrated this promising path with a virtual reality setup in which the force-feedback of a guidance system is adapted in real-time to the "mental workload" of the user [23]. A first application of this approach is a medical simulator in which virtual assistances are automatically adapted to surgeon and trainee's mental activity as illustrated Figure 9 . Such results pave the way to future virtual reality systems which would automatically reconfigure and adapt to cerebral inputs and cognitive processes.

6.4. Collaborative Virtual Environments

6.4.1. Collaborative exploration in multi-scale shared virtual environments

Participants: Thierry Duval [contact], Thi Thuong Huyen Nguyen.

Exploration of large-scale 3D Virtual Environments (VEs) is often difficult because of lack of familiarity with complex virtual worlds, lack of spatial information that can be offered to users and lack of sensory details compared to the exploration of real environments. To address this problem, we presented a set of metaphors for assisting users in collaborative navigation to perform common exploration tasks in shared collaborative virtual environments [38], [56]. Our propositions consist in three guiding techniques in the form of navigation aids to enable one or several users to help one main user (exploring user) to explore the VE efficiently. These three techniques consist in drawing directional arrows, lighting up path to follow, and orienting a compass to show a direction to the exploring user. Our experimental results could show that although the directional arrows and compass surpassed the light source in a navigation task, these three techniques are completely appropriate for guiding a user in 3D complex VEs.

Figure 9. Medical simulator adapted to a BCI output. The user manipulates a virtual needle and has to insert it into a virtual liver to reach a tumor. Visual and haptic assistances are activated when a high mental workload is detected which corresponds to a more difficult manipulation of the needle.

6.4.2. Improving the awareness of collaboration in 3D virtual environments

Participants: Thierry Duval [contact], Thi Thuong Huyen Nguyen, Valérie Gouranton.

When a user is fully immersed within a Virtual Environments (VE) through a large immersive display system, his feeling of presence can be altered because of disturbing interactions with his physical environment, such as collision with hardware parts of the system or loss of tracking. This alteration can be avoided by taking into account the physical features of the user and to embed them in the VE. In [19] we could present how we use the Immersive Interactive Virtual Cabin (IIVC) model to obtain such a virtual representation of the physical environment of the user and we illustrated how it can be used to guide efficiently a user for a navigation task in a VE. We also presented how we can add 3D representations of 2D interaction tools in order to cope with asymmetrical collaborative configurations, providing 3D cues for users in order to understand the actions of the other users even if they are not fully immersed in the shared virtual environment. Last, we explained how we could enhance 3D interaction and collaboration by embedding a symbolic 3D representation of the user that would give 3D information about his posture.

6.4.3. Sharing and bridging information: application to ergonomics

Participant: Thierry Duval [contact].

We introduced a collaborative virtual environment usable to conduct ergonomic design sessions, involving the worker, ergonomists and engineers [40]. We focused particularly on the representation of the ergonomic evaluation and the interaction between an ergonomist and the main user (worker). An ergonomic evaluation of the postures was presented. An interaction architecture between the main user and an ergonomist based on the combination of animation modes of two linked manikins was also proposed. Preliminary results and future developments of the CVE (e.g. additional ergonomic evaluation tools, graphical enhancement, interaction enhancement) were then presented.

6.4.4. User embodiment and collaboration in virtual environments for training

Participants: Bruno Arnaldi, Valérie Gouranton [contact], Thomas Lopez, Florian Nouviale, Rozenn Bouville Berthelot.

In Collaborative Virtual Environments for Training (CVET), a group can learn and practice the completion of a task as a team using all the assets provided by Virtual Reality. We presented a novel mechanism that allows real and virtual humans to dynamically exchange the control of their embodiment in virtual environments [41]. Such a mechanism raises two important issues: the possibility of dynamic embodiment exchanges between real humans and virtual humans and the continuity of actions of the team members after an exchange. To address these issues we introduce a new entity, the Perceptive Puppet that abstracts real and virtual humans into one common entity containing its own knowledge.

In addition, in CVET different roles need to be played by actors, i.e. virtual agents or users. In order to abstract an actor from its embodiment in the virtual world, we have introduced a new entity, the Shell [36]. Through the Shell, users and virtual agents are able to collaborate in the same manner during the training. In addition to the embodiment's control, the Shell gathers and carries knowledge and provides interaction inputs. This knowledge and those inputs can be accessed and used homogeneously by both users and virtual agents to help them to perform the procedure.

Hycomes Team

5. New Results

5.1. Hybrid Systems Modeling

Participants: Albert Benveniste, Benoît Caillaud.

5.1.1. Type-Based Analysis of Causality Loops In Hybrid Systems Modelers

Explicit hybrid systems modelers like Simulink / Stateflow allow for programming both discrete- and continuous-time behaviors with complex interactions between them. A key issue in their compilation is the static detection of algebraic or causality loops. Such loops can cause simulations to deadlock and prevent the generation of statically scheduled code. We have addressed this issue for a hybrid modeling language that combines synchronous Lustre-like data-flow equations with Ordinary Differential Equations (ODEs) [6], [9]. We introduce the operator last(x) for the left-limit of a signal x. This operator is used to break causality loops and permits a uniform treatment of discrete and continuous state variables. The semantics relies on non-standard analysis, defining an execution as a sequence of infinitesimally small steps. A signal is deemed causally correct when it can be computed sequentially and only progresses by infinitesimal steps outside of discrete events. The causality analysis takes the form of a simple type system. In well-typed programs, signals are proved continuous during integration and can be translated into sequential code for integration with off-the-shelf ODE solvers. The effectiveness of this system is illustrated with several examples written in Zélus [9], a Lustre-like synchronous language extended with hierarchical automata and ODEs.

5.1.2. Semantics of multi-mode DAE systems

Hybrid systems modelers exhibit a number of difficulties related to the mix of continuous and discrete dynamics and sensitivity to the discretization scheme. Modular modeling, where subsystems models can be simply assembled with no rework, calls for using Differential Algebraic Equations (DAE). In turn, DAE are strictly more difficult than ODE. They require sophisticated pre-processing using various notions of index before they can be submitted to a solver. We have studied some fundamental issues raised by the modeling and simulation of hybrid systems involving DAEs [10]. The objective of this work is to serve for the evolution and the design of future releases of the Modelica language for such systems. We focus on the following questions:

- What is the proper notion of index for a hybrid DAE system?
- What are the primitive statements needed for a DAE hybrid systems modeler?

The differentiation index for DAE explicitly relies on everything being differentiable. Therefore, generalizations to hybrid systems must be done with caution. We propose to rely on non-standard analysis for this. Non-standard analysis formalizes differential equations as discrete step transition systems with infinitesimal time basis. We can thus bring hybrid DAE systems to their non-standard form, where the notion of difference index can be firmly used. From this study, general hints for future releases of Modelica can be drawn.

5.2. Surgical Process Mining with Test and Flip Net Synthesis

Participant: Benoît Caillaud.

Surgical process modeling aims at providing an explicit representation of surgical procedural knowledge. *Surgical process models* are inferred from a set of surgical procedure recordings, and represent in a concise manner concurrency, causality and conflict relations between actions. In the context of the S3PM project (Section 6.1), we have investigated the use of *test and flip* nets, a mild extension of flip-flop nets, to represent surgical process models. A test and flip net synthesis algorithm, based on linear algebraic methods in the $Z/2Z$ ring is detailed. Experimental results regarding the use of this synthesis algorithm to automate the construction of simple surgical process models are also presented.

[9] http://zelus.di.ens.fr

I4S Project-Team

5. New Results

5.1. identification of linear systems

5.1.1. Evaluation of confidence intervals and computation of sensitivities for subspace methods

Participants: Michael Doehler, Laurent Mevel.

Stochastic Subspace Identification methods have been extensively used for the modal analysis of mechanical, civil or aeronautical structures for the last ten years. So-called stabilization diagrams are used, where modal parameters are estimated at successive model orders, leading to a graphical procedure where the physical modes of the system are extracted and separated from spurious modes. Recently an uncertainty computation scheme has been derived allowing the computation of uncertainty bounds for modal parameters at some given model order. In this paper, two problems are addressed. Firstly, a fast computation scheme is proposed reducing the computational burden of the uncertainty computation scheme by an order of magnitude in the model order compared to a direct implementation. Secondly, a new algorithm is proposed to derive the uncertainty bounds for the estimated modes at all model orders in the stabilization diagram. It is shown that this new algorithm is both computationally and memory efficient, reducing the computational burden by two orders of magnitude in the model order[14].

5.1.2. Subspace methods in frequency domain

Participants: Philippe Mellinger, Michael Doehler, Laurent Mevel.

In this paper a combined subspace algorithm and a way to quantify uncertainties of its resulting identified modal parameter has been presented. Even if the algorithm is data-driven, it was proven that uncertainties can still be quantified by using the square subspace matrix without any modification neither on the identified modal parameters or on the stabilization diagrams. A comparison between uncertainty quantification based on this data-driven combined subspace algorithm and the well-known covariance-driven stochastic subspace algorithm shows good results on this new method. Both values and confidence intervals are similar. However combined algorithm gives better results considering spurious modes. [27].

5.1.3. Subspace Identification for Linear Periodically Time-varying Systems

Participants: Laurent Mevel, Ahmed Jhinaoui.

Many systems such as turbo-generators, wind turbines and helicopters show intrinsic time-periodic behaviors. Usually, these structures are considered to be faithfully modeled as Linear Time-Invariant (LTI). In some cases where the rotor is anisotropic, this modeling does not hold and the equations of motion lead necessarily to a Linear Periodically Time-Varying (referred to as LPTV in the control and digital signal field or LTP in the mechanical and nonlinear dynamics world) model. Classical modal analysis methodologies based on the classical time-invariant eigenstructure (frequencies and damping ratios) of the system no more apply. This is the case in particular for subspace methods. For such time-periodic systems, the modal analysis can be described by characteristic exponents called Floquet multipliers. The aim of this paper is to suggest a new subspace-based algorithm that is able to extract these multipliers and the corresponding frequencies and damping ratios. The algorithm is then tested on a numerical model of a hinged-bladed helicopter on the ground. [22], [23], [18].

5.2. damage detection for mechanical structures

5.2.1. Damage detection and localisation

Participants: Michael Doehler, Luciano Gallegos, Laurent Mevel.

Mechanical systems under vibration excitation are prime candidate for being modeled by linear time invariant systems. Damage detection in such systems relates to the monitoring of the changes in the eigenstructure of the corresponding linear system, and thus reflects changes in modal parameters (frequencies, damping, mode shapes) and finally in the finite element model of the structure. Damage localization using both finite element information and modal parameters estimated from ambient vibration data collected from sensors is possible by the Stochastic Dynamic Damage Location Vector (SDDLV) approach. Damage is related to some residual derived from the kernel of the difference between transfer matrices in both reference and damage states and a model of the reference state. Deciding that this residual is zero is up to now done using an empirically defined threshold. In this paper, we show how the uncertainty in the estimates of the state space system can be used to derive uncertainty bounds on the damage localization residuals to decide about the damage location with a hypothesis test.[13], [21], [26].

5.2.2. Robust subspace damage detection

Participants: Michael Doehler, Laurent Mevel.

The detection of changes in the eigenstructure of a linear time invariant system by means of a subspace-based residual function has been proposed previously. While enjoying some success in its applicability in particular in the context of vibration monitoring, the robustness of this framework against changes in the noise properties has not been properly addressed yet. In this paper, a new robust residual is proposed and the robustness of its statistics against changes in the noise covariances is shown. The complete theory for hypothesis testing for fault detection is derived and a numerical illustration is provided[16].

5.2.2.1. Feasibility of reflectometry techniques for non destructive evaluation of external post-tensioned cables

Participant: Qinghua Zhang.

Nowadays a considerable number of bridges is reaching an age when renovating operations become necessary. For some bridges, external post-tension is realized with cables protected in ducts, with the residual internal space imperfectly filled with a fluid cement grout. Detecting the problems of injection in the ducts is visually impossible from the outside. Through a collaboration with the SISYPHE project-team, the feasibility of reflectometry techniques for cable health monitoring is investigated via numerical simulations and laboratory experiments. The main idea consists in adding electrically conductive tapes along a duct so that the duct and the added tapes can be treated as an electrical transmission line. It is then possible to apply advanced reflectometry methods developed by the SISYPHE project-team, initially for true electric cables.

IPSO Project-Team

5. New Results

5.1. Multi-revolution composition methods for highly oscillatory differential equations

In [45], we introduce a new class of multi-revolution composition methods (MRCM) for the approximation of the Nth-iterate of a given near-identity map. When applied to the numerical integration of highly oscillatory systems of differential equations, the technique benefits from the properties of standard composition methods: it is intrinsically geometric and well-suited for Hamiltonian or divergence-free equations for instance. We prove error estimates with error constants that are independent of the oscillatory frequency. Numerical experiments, in particular for the nonlinear Schrödinger equation, illustrate the theoretical results, as well as the efficiency and versatility of the methods.

5.2. Weak second order multi-revolution composition methods for highly oscillatory stochastic differential equations with additive or multiplicative noise

In [61], we introduce a class of numerical methods for highly oscillatory systems of stochastic differential equations with general noncommutative noise. We prove global weak error bounds of order two uniformly with respect to the stiffness of the oscillations, which permits to use large time steps. The approach is based on the micro-macro framework of multi-revolution composition methods recently introduced for deterministic problems and inherits its geometric features, in particular to design integrators preserving exactly quadratic first integral. Numerical experiments, including the stochastic nonlinear Schrödinger equation with space-time multiplicative noise, illustrate the performance and versatility of the approach.

5.3. High order numerical approximation of the invariant measure of ergodic SDEs

In [41], we introduce new sufficient conditions for a numerical method to approximate with high order of accuracy the invariant measure of an ergodic system of stochastic differential equations, independently of the weak order of accuracy of the method. We then present a systematic procedure based on the framework of modified differential equations for the construction of stochastic integrators that capture the invariant measure of a wide class of ergodic SDEs (Brownian and Langevin dynamics) with an accuracy independent of the weak order of the underlying method. Numerical experiments confirm our theoretical findings.

5.4. PIROCK: a swiss-knife partitioned implicit-explicit orthogonal Runge-Kutta Chebyshev integrator for stiff diffusion-advection-reaction problems with or without noise

In [13], a partitioned implicit-explicit orthogonal Runge-Kutta method (PIROCK) is proposed for the time integration of diffusion-advection-reaction problems with possibly severely stiff reaction terms and stiff stochastic terms. The diffusion terms are solved by the explicit second order orthogonal Chebyshev method (ROCK2), while the stiff reaction terms (solved implicitly) and the advection and noise terms (solved explicitly) are integrated in the algorithm as finishing procedures. It is shown that the various coupling (between diffusion, reaction, advection and noise) can be stabilized in the PIROCK method. The method, implemented in a single black-box code that is fully adaptive, provides error estimators for the various terms present in the problem, and requires from the user solely the right-hand side of the differential equation. Numerical experiments and comparisons with existing Chebyshev methods, IMEX methods and partitioned methods show the efficiency and flexibility of our new algorithm.

5.5. An offline-online homogenization strategy to solve quasilinear two-scale problems at the cost of one-scale problems

In [39], inspired by recent analyses of the finite element heterogeneous multiscale method and the reduced basis technique for nonlinear problems, we present a simple and concise finite element algorithm for the reliable and efficient resolution of elliptic or parabolic multiscale problems of nonmonotone type. Solutions of appropriate cell problems on sampling domains are selected by a greedy algorithm in an offline stage and assembled in a reduced basis (RB). This RB is then used in an online stage to solve two-scale problems at a computational cost comparable to the single-scale case. Both the offline and the online cost are independent of the smallest scale in the physical problem. The performance and accuracy of the algorithm are illustrated on 2D and 3D stationary and evolutionary nonlinear multiscale problems.

5.6. Reduced basis finite element heterogeneous multiscale method for quasilinear elliptic homogenization problems

In [40], a reduced basis nite element heterogeneous multiscale method (RB-FE-HMM) for a class of nonlinear homogenization elliptic problems of nonmonotone type is introduced. In this approach, the solutions of the micro problems needed to estimate the macroscopic data of the homogenized problem are selected by a Greedy algorithm and computed in an online stage. It is shown that the use of reduced basis (RB) for nonlinear numerical homogenization reduces considerably the computational cost of the nite element heterogeneous multiscale method (FE-HMM). As the precomputed microscopic functions depend nonlinearly on the macroscopic solution, we introduce a new a posteriori error estimator for the Greedy algorithm that guarantees the convergence of the online Newton method. A priori error estimates and uniqueness of the numerical solution are also established. Numerical experiments illustrate the e ciency of the proposed method.

5.7. Weak second order explicit stabilized methods for stiff stochastic differential equations

In [16], we introduce a new family of explicit integrators for stiff Itô stochastic differential equations (SDEs) of weak order two. These numerical methods belong to the class of one-step stabilized methods with extended stability domains and do not suffer from the stepsize reduction faced by standard explicit methods. The family is based on the standard second order orthogonal Runge-Kutta Chebyshev methods (ROCK2) for deterministic problems. The convergence, and the mean-square and asymptotic stability properties of the methods are analyzed. Numerical experiments, including applications to nonlinear SDEs and parabolic stochastic partial differential equations are presented and confirm the theoretical results.

5.8. Mean-square A-stable diagonally drift-implicit integrators of weak second order for stiff Itô stochastic differential equations

In [15], we introduce two drift-diagonally-implicit and derivative-free integrators for stiff systems of Itô stochastic differential equations with general non-commutative noise which have weak order 2 and deterministic order 2, 3, respectively. The methods are shown to be mean-square A-stable for the usual complex scalar linear test problem with multiplicative noise and improve significantly the stability properties of the drift-diagonally-implicit methods previously introduced [K. Debrabant and A. Rößler, Appl. Num. Math., 59, 2009].

5.9. Two-Scale Macro-Micro decomposition of the Vlasov equation with a strong magnetic field

In [25], we build a Two-Scale Macro-Micro decomposition of the Vlasov equation with a strong magnetic field. This consists in writing the solution of this equation as a sum of two oscillating functions with circonscribed oscillations. The first of these functions has a shape which is close to the shape of the Two-Scale limit of the solution and the second one is a correction built to offset this imposed shape. The aim of such a decomposition is to be the starting point for the construction of Two-Scale Asymptotic-Preserving Schemes.

5.10. A dynamic multi-scale model for transient radiative transfer calculations

In [33], a dynamic multi-scale model which couples the transient radiative transfer equation (RTE) and the diffusion equation (DE) is proposed and validated. It is based on a domain decomposition method where the system is divided into a mesoscopic subdomain, where the RTE is solved, and a macroscopic subdomain where the DE is solved. A buffer zone is introduced between the mesoscopic and the macroscopic subdomains, as proposed by Degond and Jin, who solve a coupled system of two equations, one at the mesoscopic and the other at the macroscopic scale. The DE and the RTE are coupled through the equations inside the buffer zone, instead of being coupled through a geometric interface like in standard domain decomposition methods. One main advantage is that no boundary or interface conditions are needed for the DE. The model is compared to Monte Carlo, finite volume and P1 solutions in one dimensional stationary and transient test cases, and presents promising results in terms of trade-off between accuracy and computational requirements.

5.11. Quasi-periodic solutions of the 2D Euler equation

In [24], we consider the two-dimensional Euler equation with periodic boundary conditions. We construct time quasi-periodic solutions of this equation made of localized travelling profiles with compact support propagating over a stationary state depending on only one variable. The direction of propagation is orthogonal to this variable, and the support is concentrated on flat strips of the stationary state. The frequencies of the solution are given by the locally constant velocities associated with the stationary state.

5.12. Optimization and parallelization of Emedge3D on shared memory architecture

In [38], a study of techniques used to speedup a scientific simulation code is presented. The techniques include sequential optimizations as well as the parallelization with OpenMP. This work is carried out on two different multicore shared memory architectures, namely a cutting edge 8x8 core CPU and a more common 2x6 core board. Our target application is representative of many memory bound codes, and the techniques we present show how to overcome the burden of the memory bandwidth limit, which is quickly reached on multi-core or many-core with shared memory architectures. To achieve efficient speedups, strategies are applied to lower the computation costs, and to maximize the use of processors caches. Optimizations are: minimizing memory accesses, simplifying and reordering computations, and tiling loops. On 12 cores processor Intel X5675, aggregation of these optimizations results in an execution time 21.6 faster, compared to the original version on one core.

5.13. Vlasov on GPU (VOG Project)

In [58], we are concerned with the numerical simulation of the Vlasov-Poisson set of equations using semi- Lagrangian methods on Graphical Processing Units (GPU). To accomplish this goal, modifications to traditional methods had to be implemented. First and foremost, a reformulation of semi-Lagrangian methods is performed, which enables us to rewrite the governing equations as a circulant matrix operating on the vector of unknowns. This product calculation can be performed efficiently using FFT routines. Second, to overcome the limitation of single precision inherent in GPU, a δf type method is adopted which only needs refinement in specialized areas of phase space but not throughout. Thus, a GPU Vlasov-Poisson solver can indeed perform high precision simulations (since it uses very high order reconstruction methods and a large number of grid points in phase space). We show results for rather academic test cases on Landau damping and also for physically relevant phenomena such as the bump on tail instability and the simulation of Kinetic Electrostatic Electron Nonlinear (KEEN) waves.

5.14. Uniformly accurate numerical schemes for highly oscillatory Klein-Gordon and nonlinear Schrödinger equations

In [37], we are interested in the numerical simulation of nonlinear Schrödinger and Klein-Gordon equations. We present a general strategy to construct numerical schemes which are uniformly accurate with respect to the oscillation frequency. This is a stronger feature than the usual so called "Asymptotic preserving" property, the last being also satisfied by our scheme in the highly oscillatory limit. Our strategy enables to simulate the oscillatory problem without using any mesh or time step refinement, and the orders of our schemes are preserved uniformly in all regimes. In other words, since our numerical method is not based on the derivation and the simulation of asymptotic models, it works in the regime where the solution does not oscillate rapidly, in the highly oscillatory limit regime, and in the intermediate regime with the same order of accuracy. The method is based on two main ingredients. First, we embed our problem in a suitable "two-scale" reformulation with the introduction of an additional variable. Then a link is made with classical strategies based on Chapman-Enskog expansions in kinetic theory despite the dispersive context of the targeted equations, allowing to separate the fast time scale from the slow one. Uniformly accurate (UA) schemes are eventually derived from this new formulation and their properties and performances are assessed both theoretically and numerically.

5.15. Asymptotic preserving schemes for the Wigner-Poisson-BGK equations in the diffusion limit

In [26], we focus on the numerical simulation of the Wigner-Poisson-BGK equation in the diffusion asymptotics. Our strategy is based on a "micro-macro" decomposition, which leads to a system of equations that couple the macroscopic evolution (diffusion) to a microscopic kinetic contribution for the fluctuations. A semi-implicit discretization provides a numerical scheme which is stable with respect to the small parameter ε (mean free path) and which possesses the following properties: (i) it enjoys the asymptotic preserving property in the diffusive limit; (ii) it recovers a standard discretization of the Wigner-Poisson equation in the collisionless regime. Numerical experiments confirm the good behaviour of the numerical scheme in both regimes. The case of a spatially dependent $\varepsilon(x)$ is also investigated.

5.16. Existence and stability of solitons for fully discrete approximations of the nonlinear Schrödinger equation

In [19], we study the long time behavior of a discrete approximation in time and space of the cubic nonlinear Schrödinger equation on the real line. More precisely, we consider a symplectic time splitting integrator applied to a discrete nonlinear Schrödinger equation with additional Dirichlet boundary conditions on a large interval. We give conditions ensuring the existence of a numerical soliton which is close in energy norm to the continuous soliton. Such result is valid under a CFL condition between the time and space stepsizes. Furthermore we prove that if the initial datum is symmetric and close to the continuous soliton, then the associated numerical solution remains close to the orbit of the continuous soliton for very long times.

5.17. Asymptotic preserving schemes for the Klein-Gordon equation in the non-relativistic limit regime

In [32], we consider the Klein-Gordon equation in the non-relativistic limit regime, i.e. the speed of light c tending to infinity. We construct an asymptotic expansion for the solution with respect to the small parameter depending on the inverse of the square of the speed of light. As the first terms of this asymptotic can easily be simulated our approach allows us to construct numerical algorithms that are robust with respect to the large parameter c producing high oscillations in the exact solution.

5.18. Sobolev stability of plane wave solutions to the cubic nonlinear Schrödinger equation on a torus

In [31], it is shown that plane wave solutions to the cubic nonlinear Schrödinger equation on a torus behave orbitally stable under generic perturbations of the initial data that are small in a high-order Sobolev norm, over long times that extend to arbitrary negative powers of the smallness parameter. The perturbation stays small in the same Sobolev norm over such long times. The proof uses a Hamiltonian reduction and transformation and, alternatively, Birkhoff normal forms or modulated Fourier expansions in time.

5.19. Weak backward error analysis for overdamped Langevin equation

In [57], we consider an overdamped Langevin stochastic differential equation and show a weak backward error analysis result for its numerical approximations defined by implicit methods. In particular, we prove that the generator associated with the numerical solution coincides with the solution of a modified Kolmogorov equation up to high order terms with respect to the stepsize. This implies that every measure of the numerical scheme is close to a modified invariant measure obtained by asymptotic expansion. Moreover, we prove that, up to negligible terms, the dynamic associated with the implicit scheme considered is exponentially mixing.

5.20. Weak backward error analysis for Langevin equation

In [56], We consider numerical approximations of stochastic Langevin equations by implicit methods. We show a weak backward error analysis result in the sense that the generator associated with the numerical solution coincides with the solution of a modified Kolmogorov equation up to high order terms with respect to the stepsize. This implies that every measure of the numerical scheme is close to a modified invariant measure obtained by asymptotic expansion. Moreover, we prove that, up to negligible terms, the dynamic associated with the implicit scheme considered is exponentially mixing.

5.21. Approximation of the invariant law of SPDEs: error analysis using a Poisson equation for a full-discretization scheme

In [44], we study the long-time behavior of fully discretized semilinear SPDEs with additive space-time white noise, which admit a unique invariant probability measure μ. We show that the average of regular enough test functions with respect to the (possibly non unique) invariant laws of the approximations are close to the corresponding quantity for μ.

More precisely, we analyze the rate of the convergence with respect to the different discretization parameters. Here we focus on the discretization in time thanks to a scheme of Euler type, and on a Finite Element discretization in space.

The results rely on the use of a Poisson equation; we obtain that the rates of convergence for the invariant laws are given by the weak order of the discretization on finite time intervals: order $1/2$ with respect to the time-step and order 1 with respect to the mesh-size.

5.22. An asymptotic preserving scheme based on a new formulation for NLS in the semiclassical limit

In [20], we consider the semiclassical limit for the nonlinear Schrodinger equation. We introduce a phase/amplitude representation given by a system similar to the hydrodynamical formulation, whose novelty consists in including some asymptotically vanishing viscosity. We prove that the system is always locally well-posed in a class of Sobolev spaces, and globally well-posed for a fixed positive Planck constant in the one-dimensional case. We propose a second order numerical scheme which is asymptotic preserving. Before singularities appear in the limiting Euler equation, we recover the quadratic physical observables as well as the wave function with mesh size and time step independent of the Planck constant. This approach is also well suited to the linear Schrodinger equation.

5.23. Asymptotic Preserving schemes for highly oscillatory Vlasov-Poisson equations

The work [28] is devoted to the numerical simulation of a Vlasov-Poisson model describing a charged particle beam under the action of a rapidly oscillating external field. We construct an Asymptotic Preserving numerical scheme for this kinetic equation in the highly oscillatory limit. This scheme enables to simulate the problem without using any time step refinement technique. Moreover, since our numerical method is not based on the derivation of the simulation of asymptotic models, it works in the regime where the solution does not oscillate rapidly, and in the highly oscillatory regime as well. Our method is based on a "two scale" reformulation of the initial equation, with the introduction of an additional periodic variable.

5.24. Uniformly accurate numerical schemes for highly oscillatory Klein-Gordon and nonlinear Schrödinger equations

The work [37] is devoted to the numerical simulation of nonlinear Schrödinger and Klein-Gordon equations. We present a general strategy to construct numerical schemes which are uniformly accurate with respect to the oscillation frequency. This is a stronger feature than the usual so called ”Asymptotic preserving" property, the last being also satisfied by our scheme in the highly oscillatory limit. Our strategy enables to simulate the oscillatory problem without using any mesh or time step refinement, and the orders of our schemes are preserved uniformly in all regimes. In other words, since our numerical method is not based on the derivation and the simulation of asymptotic models, it works in the regime where the solution does not oscillate rapidly, in the highly oscillatory limit regime, and in the intermediate regime with the same order of accuracy. In the same spirit as in [28], the method is based on two main ingredients. First, we embed our problem in a suitable ”two-scale" reformulation with the introduction of an additional variable. Then a link is made with classical strategies based on Chapman-Enskog expansions in kinetic theory despite the dispersive context of the targeted equations, allowing to separate the fast time scale from the slow one. Uniformly accurate (UA) schemes are eventually derived from this new formulation and their properties and performances are assessed both theoretically and numerically.

5.25. On the controllability of quantum transport in an electronic nanostructure

In [59], we investigate the controllability of quantum electrons trapped in a two-dimensional device, typically a MOS field-effect transistor. The problem is modeled by the Schrödinger equation in a bounded domain coupled to the Poisson equation for the electrical potential. The controller acts on the system through the boundary condition on the potential, on a part of the boundary modeling the gate. We prove that, generically with respect to the shape of the domain and boundary conditions on the gate, the device is controllable. We also consider control properties of a more realistic nonlinear version of the device, taking into account the self-consistent electrostatic Poisson potential.

5.26. The Interaction Picture method for solving the generalized nonlinear Schrödinger equation in optics

The ”interaction picture” (IP) method is a very promising alternative to Split-Step methods for solving certain type of partial differential equations such as the nonlinear Schrödinger equation involved in the simulation of wave propagation in optical fibers. The method exhibits interesting convergence properties and is likely to provide more accurate numerical results than cost comparable Split-Step methods such as the Symmetric Split-Step method. In [42] we investigate in detail the numerical properties of the IP method and carry out a precise comparison between the IP method and the Symmetric Split-Step method.

5.27. Solving highly-oscillatory NLS with SAM: numerical efficiency and geometric properties

In [46], we present the Stroboscopic Averaging Method (SAM), recently introduced in [7,8,10,12], which aims at numerically solving highly-oscillatory differential equations. More specifically, we first apply SAM to the Schrödinger equation on the 1-dimensional torus and on the real line with harmonic potential, with the aim of assessing its efficiency: as compared to the well-established standard splitting schemes, the stiffer the problem is, the larger the speed-up grows (up to a factor 100 in our tests). The geometric properties of SAM are also explored: on very long time intervals, symmetric implementations of the method show a very good preservation of the mass invariant and of the energy. In a second series of experiments on 2-dimensional equations, we demonstrate the ability of SAM to capture qualitatively the long-time evolution of the solution (without spurring high oscillations).

5.28. Analysis of models for quantum transport of electrons in graphene layers

In [51], we present and analyze two mathematical models for the self consistent quantum transport of electrons in a graphene layer. We treat two situations. First, when the particles can move in all the plane R^2, the model takes the form of a system of massless Dirac equations coupled together by a selfconsistent potential, which is the trace in the plane of the graphene of the 3D Poisson potential associated to surface densities. In this case, we prove local in time existence and uniqueness of a solution in $H^s(R^2)$, for $s > 3/8$ which includes in particular the energy space $H^{1/2}(R^2)$. The main tools that enable to reach $s \in (3/8, 1/2)$ are the dispersive Strichartz estimates that we generalized here for mixed quantum states. Second, we consider a situation where the particles are constrained in a regular bounded domain Ω. In order to take into account Dirichlet boundary conditions which are not compatible with the Dirac Hamiltonian H_0, we propose a different model built on a modified Hamiltonian displaying the same energy band diagram as H_0 near the Dirac points. The well-posedness of the system in this case is proved in H^s_A, the domain of the fractional order Dirichlet Laplacian operator, for $1/2 \leq s$.

5.29. Analysis of a large number of Markov chains competing for transitions

In [18], we consider the behavior of a stochastic system composed of several identically distributed, but non independent, discrete-time absorbing Markov chains competing at each instant for a transition. The competition consists in determining at each instant, using a given probability distribution, the only Markov chain allowed to make a transition. We analyze the first time at which one of the Markov chains reaches its absorbing state. When the number of Markov chains goes to infinity, we analyze the asymptotic behavior of the system for an arbitrary probability mass function governing the competition. We give conditions for the existence of the asymptotic distribution and we show how these results apply to cluster-based distributed systems when the competition between the Markov chains is handled by using a geometric distribution.

5.30. Markov Chains Competing for Transitions: Application to Large-Scale Distributed Systems

In [17], we consider the behavior of a stochastic system composed of several identically distributed, but non independent, discrete-time absorbing Markov chains competing at each instant for a transition. The competition consists in determining at each instant, using a given probability distribution, the only Markov chain allowed to make a transition. We analyze the first time at which one of the Markov chains reaches its absorbing state. We obtain its distribution and its expectation and we propose an algorithm to compute these quantities. We also exhibit the asymptotic behavior of the system when the number of Markov chains goes to infinity. Actually, this problem comes from the analysis of large-scale distributed systems and we show how our results apply to this domain.

5.31. Existence of densities for the 3D Navier–Stokes equations driven by Gaussian noise

In [30], we prove three results on the existence of densities for the laws of finite dimensional functionals of the solutions of the stochastic Navier-Stokes equations in dimension 3. In particular, under very mild assumptions on the noise, we prove that finite dimensional projections of the solutions have densi- ties with respect to the Lebesgue measure which have some smoothness when measured in a Besov space. This is proved thanks to a new argument inspired by an idea introduced by N. Fournier and J. Printems.

5.32. Invariant measure of scalar first-order conservation laws with stochastic forcing

In [50], we assume an hypothesis of non-degeneracy of the flux and study the long-time behaviour of periodic scalar first-order conservation laws with stochastic forcing in any space dimension. For sub-cubic fluxes, we show the existence of an invariant measure. Moreover for sub-quadratic fluxes we show uniqueness and ergodicity of the invariant measure. Also, since this invariant measure is supported by L^p for some p small, we are led to generalize to the stochastic case the theory of L^1 solutions developed by Chen and Perthame.

5.33. Degenerate Parabolic Stochastic Partial Differential Equations: Quasilinear case

In [49], we study the Cauchy problem for a quasilinear degenerate parabolic stochastic partial differential equation driven by a cylindrical Wiener process. In particular, we adapt the notion of kinetic formulation and kinetic solution and develop a well-posedness theory that includes also an L^1-contraction property. In comparison to the previous works of the authors concerning stochastic hyperbolic conservation laws and semilinear degenerate parabolic SPDEs, the present result contains two new ingredients that provide simpler and more effective method of the proof: a generalized Itô formula that permits a rigorous derivation of the kinetic formulation even in the case of weak solutions of certain nondegenerate approximations and a direct proof of strong convergence of these approximations to the desired kinetic solution of the degenerate problem.

5.34. Existence of densities for stable-like driven SDE's with Hölder continuous coefficients

In [29], we consider a multidimensional stochastic differential equation driven by a stable-like Lévy process. We prove that the law of the solution immediately has a density in some Besov space, under some non-degeneracy condition on the driving Lévy process and some very light Hölder-continuity assumptions on the drift and diffusion coefficients.

5.35. Ergodicity results for the stochastic Navier-Stokes equations: an introduction

In the chapter [36], we give an overview of the results on ergodicity for the stochastic Navier-Stokes equations. We first explain the basis on SPDEs and on the concept of invariant measures and ergodicity. Then, in the 2D case, we introduce progressively the various methods, finishing with a celebrated result due to M. Hairer and J. Mattingly on ergodicity with very degenerated noises. In the 3D case, the theory is much less complete. Nonetheles, we show that it is possible to construct Markov evolutions and, under some non degenary assumptions on the noise, to obtain ergodicity.

5.36. Weak truncation error estimates for elliptic PDEs with lognormal coefficients

In [22], we are interested in the weak error committed on the solution of an elliptic partial differential equation with a lognormal coefficient, resulting from the approximation of the lognormal coefficient through a Karhunen-Loéve expansion. We improve results of a previous work, in which L^p-estimates of the weak error are provided. Only small enough values of p (the corresponding values of p depend on the space dimension) could be considered and such bounds are not sufficient to be applied to practical cases. Moreover, the optimality of this weak order (which turns out to be twice the strong order) has not been studied numerically. Therefore, the aim of this paper is double. First we improve drastically the weak error estimate by providing a bound of the C^1-norm of the weak error. This requires regularity results in Hölder spaces, with explicit bounds for the constants. We also consider much more general test functions in the definition of the weak error. Finally, we show the optimality of the weak order and illustrate this weak convergence with numerical results.

5.37. Optimized high-order splitting methods for some classes of parabolic equations

In [21], we are concerned with the numerical solution obtained by splitting methods of certain parabolic partial differential equations. Splitting schemes of order higher than two with real coefficients necessarily involve negative coefficients. It has been demonstrated that this second-order barrier can be overcome by using splitting methods with complex-valued coefficients (with positive real parts). In this way, methods of orders 3 to 14 by using the Suzuki-Yoshida triple (and quadruple) jump composition procedure have been explicitly built. Here we reconsider this technique and show that it is inherently bounded to order 14 and clearly sub-optimal with respect to error constants. As an alternative, we solve directly the algebraic equations arising from the order conditions and construct methods of orders 6 and 8 that are the most accurate ones available at present time, even when low accuracies are desired. We also show that, in the general case, 14 is not an order barrier for splitting methods with complex coefficients with positive real part by building explicitly a method of order 16 as a composition of methods of order 8.

5.38. Higher-Order Averaging, Formal Series and Numerical Integration III: Error Bounds

In earlier papers, it has been shown how formal series like those used nowadays to investigate the properties of numerical integrators may be used to construct high-order averaged systems or formal first integrals of Hamiltonian problems. With the new approach the averaged system (or the formal first integral) may be written down immediately in terms of (i) suitable basis functions and (ii) scalar coefficients that are computed via simple recursions. In [23], we show how the coefficients/basis functions approach may be used advantageously to derive exponentially small error bounds for averaged systems and approximate first integrals.

KERDATA Project-Team

6. New Results

6.1. A-Brain and TomusBlobs

6.1.1. Experiments with TomusBlobs at large scale

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Joint genetic and neuro-imaging data analysis may help identifying risk factors in target populations. Performing such studies on a large number of subjects is challenging as genotyping DNA chips can record several hundred thousands values per subject, while the fMRI images may contain 100k–1M volumetric picture elements. Determining statistically significant links between the two sets of data entails a massive amount of computation as one needs not only to compare all pair-wise relations but also to correct for family-wise multiple comparisons. These false positives are controlled by generating permutations of the input data set. The A-Brain initiative is such a data analysis application involving large cohorts of subjects and used to study and understand the variability that exists between individuals. Supposing that such an application could be executed on a single machine, the computation would take years. Cloud infrastructures have the potential to decrease this computation time to days, by parallellizing and scaling out the application. In order to execute this computation in parallel at a large scale, we noticed that the A-Brain application can be easily described as a MapReduce process. The problem was further divided into 28,000 computation tasks, which were executed as map jobs.

The experiment timespan was 14 days and was performed across 4 cloud deployments in 2 different US Azure data centers — North and West. The processing time for a map job is approximatively 2 hours and there are no notable time differences between the map execution time with respect to the geographical location. This is achieved due to the load balancing of the workload, the data locality within the deployments and to the geographical partition. The global result was aggregated using a MapIterativeReduce technique which was composed of 563 reduce jobs. This reduction process eliminates the implicit barrier between mappers and reducers, the reduction process happens in parallel with the map computation. During the period of the experiment the Azure services became temporary inaccessible, due to a failure of a secured certificate. Despite this problem, the framework was capable to handle the failure due to a safety mechanism that we implemented which suspended the computation until all Azure services became available again. Regarding the lost map/reduce enqueued jobs, the monitor mechanism, which supervises the computation progress, was able to restore them. The cost of the experiment was approximatively 210,000 compute hours, where 1 compute hour is equivalent to 1 CPU running for one hour. The monetary cost of the experiment adds up to almost 20,000 $. The total amount combines the cost of the compute resources, for which a value of 0.09 $/h was considered, the persistent Azure storage cost and the outbound traffic from the data centers. As a result of this experiment, we have confirmed that brain activation signals are a heritable feature.

6.1.2. Using dedicated compute nodes for data management on public clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

A large spectrum of scientific applications, some generating data volumes exceeding petabytes, are currently being ported on clouds to build on their inherent elasticity and scalability. One of the critical needs in order to deal with this "data deluge" is an efficient, scalable and reliable storage. However, the storage services proposed by cloud providers suffer from high latencies, trading performance for availability. One alternative is to federate the local virtual disks on the compute nodes into a globally shared storage used for large intermediate or checkpoint data. This collocated storage supports a high throughput but it can be very intrusive and subject to failures that can stop the host node and degrade the application performance.

To deal with these limitations we proposed DataSteward [25], a data management system that provides a higher degree of reliability while remaining non-intrusive through the use of dedicated compute nodes. DataSteward harnesses the storage space of a set of dedicated VMs, selected using a topology-aware clustering algorithm, and has a lifetime dependent on the deployment lifetime. To capitalize on this separation, we introduced a set of scientific data processing services on top of the storage layer, that can overlap with the executing applications. We performed extensive experimentations on hundreds of cores in the Azure cloud: compared to state-of-the-art node selection algorithms, we show up to a 20 % higher throughput, which improves the overall performance of a real life scientific application by up to 45 %.

6.1.3. File transfers for workflows

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Scientific workflows typically communicate data between tasks using files. Currently, on public clouds, this is achieved by using the cloud storage services, which are unable to exploit the workflow semantics and are subject to low throughput and high latencies. To overcome these limitations, we propose in [26] an alternative leveraging data locality through direct file transfers between the compute nodes. We rely on the observation that workflows generate a set of common data access patterns that our solution exploits in conjunction with context information to self-adapt, choose the most adequate transfer protocol and expose the data layout within the virtual machines to the workflow engines. This file management system was integrated within the Microsoft Generic Worker workflow engine and was validated using synthetic benchmarks and a real-life application on the Azure cloud. The results show it can bring significant performance gains: up to 5x file transfer speedup compared to solutions based on standard cloud storage and over 25 % application timespan reduction compared to Hadoop on Azure. This work was done in colaboration with Goetz Brasche and Ramin Rezai Rad from *Microsoft Advance Technology Lab Europe*.

6.2. Optimizing MapReduce Processing

6.2.1. Optimizing MapReduce in virtualized environments

Participant: Shadi Ibrahim.

As data-intensive applications become popular in the cloud, their performance on the virtualized platform calls for empirical evaluations and technical innovations. Virtualization has become a prominent tool in data centers and is extensively leveraged in cloud environments: it enables multiple virtual machines (VMs) — with multiple operating systems and applications — to run within a physical server. However, virtualization introduces the challenging issue of providing effective QoS to VMs and preserving the high disk utilization (i.e., reducing the seek delay and rotation overhead) when allocating disk resources to VMs. We addressed these challenges by developing two Disk I/O scheduling frameworks: *Flubber* and *Pregather*.

In [17], we developed a two-level scheduling framework that decouples throughput and latency allocation to provide QoS guarantees to VMs while maintaining high disk utilization. The high-level throughput control regulates the pending requests from the VMs with an adaptive credit-rate controller, in order to meet the throughput requirements of different VMs and ensure performance isolation. Meanwhile, the low-level latency control, by the virtue of the batch and delay earliest deadline first mechanism (BD-EDF), re-orders all pending requests from VMs based on their deadlines, and batches them to disk devices taking into account the locality of accesses across VMs.

In [24], we developed a novel disk I/O scheduling framework, named *Pregather*, to improve disk I/O efficiency through exposure and exploitation of the special spatial locality in the virtualized environment (regional and sub-regional spatial locality corresponds to the virtual disk space and applications' access patterns, respectively), thereby improving the performance of disk-intensive applications (e.g., MapReduce applications) without harming the transparency feature of virtualization (without a priori knowledge of the applications' access patterns). The key idea behind Pregather is to implement an intelligent model to predict the access regularity of sub-regional spatial locality for each VM.

We evaluated *Pregather* through extensive experiments that involve multiple simultaneous applications of both synthetic benchmarks and a MapReduce application (i.e., distributed sort) on Xen-based platforms. Our experiments indicate that *Pregather* results in high disk spatial locality, yields a significant improvement in disk throughput, and ends up with improved Hadoop performance. This work was done in collaboration with Hai Jin, Song Wu and Xiao Ling from Huazhong University of Science and Technology (HUST).

6.2.2. Investigating energy efficiency in MapReduce

Participants: Shadi Ibrahim, Houssem-Eddine Chihoub, Gabriel Antoniu, Luc Bougé.

A MapReduce system spans over a multitude of computing nodes that are frequency- and voltage-scalable. Furthermore, many MapReduce applications show significant variation in CPU load during their running time. Thus, there is a significant potential for energy saving by scaling down the CPU frequency. Some power-aware data-layout techniques have been proposed to save power, at the cost of a weaker performance. MapReduce applications range from CPU-Intensive to I/O-Intensive. More importantly, a typical MapReduce application comprises many subtasks, each subtask's workload being either a computation, a disk request or a bandwidth request. As a result, there is a high potential for power reduction by scaling down the CPU when the peak CPU performance is not used.

In this ongoing work, a series of experiments are conducted to explore the implications of *Dynamic Voltage Frequency scaling* (DVFS) settings on power consumption in Hadoop-clusters, by benefitting from the current maturity in DVFS research and of the introduction of governors (e.g., *performance*, *powersave*, *ondemand*, *conservative* and *userspace*). By adapting existing DVFS governors in Hadoop clusters, we observe significant variation in performance and power consumption of the cluster with different applications when applying these governors: the different DVFS settings are only sub-optimal for different MapReduce applications. Furthermore, our results reveal that current CPU governors do not exactly reflect their design goal and may even become ineffective to manage the power consumption. Based on this analysis, we are investigating a new approach to reduce the energy consumption in Hadoop through adaptively tuning the governors and/or the CPU frequencies during the execution of MapReduce jobs.

6.2.3. Hybrid infrastructures

Participants: Alexandru Costan, Ana-Ruxandra Ion, Gabriel Antoniu.

As Map-Reduce emerges as a leading programming paradigm for data-intensive computing, today's frameworks which support it still have substantial shortcomings that limit its potential scalability. At the core of Map-Reduce frameworks lies a key component with a huge impact on their performance: the storage layer. To enable scalable parallel data processing, this layer must meet a series of specific requirements. An important challenge regards the target execution infrastructures. While the Map-Reduce programming model has become very visible in the cloud computing area, it is also subject to active research efforts on other kinds of large-scale infrastructures, such as desktop grids. We claim that it is worth investigating how such efforts (currently done in parallel) could converge, in a context where large-scale distributed platforms become more and more connected together.

We investigated several directions where there is room for such progress: they concern storage efficiency under massive-data access concurrency, scheduling, volatility and fault-tolerance. We placed our discussion in the perspective of the current evolution towards an increasing integration of large-scale distributed platforms (clouds, cloud federations, enterprise desktop grids, etc.). We proposed an approach which aims to overcome the current limitations of existing Map-Reduce frameworks, in order to achieve scalable, concurrency-optimized, fault-tolerant Map-Reduce data processing on hybrid infrastructures. We are designing and implementing our approach through an original architecture for scalable data processing: it combines two approaches, BlobSeer and BitDew, which have shown their benefits separately (on clouds and desktop grids respectively) into a unified system. The global goal is to improve the behavior of Map-Reduce-based applications on the target large-scale infrastructures. The internship of Ana-Ruxandra Ion was dedicated to this topic and showed that for reliable hybrid Map-Reduce processing, one needs to first rely on public/private cloud resources, and then to scale them up using the local, yet volatile, desktop grid resources.

6.2.4. Key partitioning techniques

Participants: Shadi Ibrahim, Gabriel Antoniu.

Data locality is a key feature in MapReduce that is extensively leveraged in data-intensive cloud systems: it avoids network saturation when processing large amounts of data by co-allocating computation and data storage, particularly for the map phase. However, our studies with Hadoop, a widely used MapReduce implementation, demonstrate that the presence of partitioning skew (partitioning skew refers to the case when a variation in either the intermediate keys' frequencies or their distributions or both among different data nodes) causes a huge amount of data transfer during the shuffle phase and leads to significant unfairness on the reduce input among different data nodes. As a result, the applications suffer from severe performance degradation due to the long data transfer during the shuffle phase along with the computation skew, particularly in reduce phase. We addressed these problems by developing a new key/value partitioning called *LEEN*.

In [16], we develop a novel algorithm named *LEEN* for locality-aware and fairness-aware key partitioning in MapReduce. *LEEN* aims at saving the network bandwidth dissipation during the shuffle phase of the MapReduce job along with balancing the reducers' inputs. *LEEN* is conducive to improve the data locality of the MapReduce execution efficiency by the virtue of the asynchronous map and reduce scheme, thereby having more control on the keys distribution in each data node. *LEEN* keeps track of the frequencies of buffered keys hosted by each data node. In doing so, *LEEN* efficiently moves buffered intermediate keys to the destination considering the location of the high frequencies along with fair distribution of reducers' inputs.

To quantify the locality, data distribution and performance of *LEEN*, we conducted a comprehensive performance evaluation study using *LEEN* in Hadoop. Our experimental results demonstrate that *LEEN* interestingly can efficiently achieve higher locality, and balance data distribution after the shuffle phase. This work was done in collaboration with Hai Jin, Song Wu and Lu Lu from Huazhong University of Science and Technology (HUST) and Bingsheng He from Nanyang Technological University (NTU).

6.3. Cloud Storage Trade-Offs: Consistency and Self-Adaptiveness

6.3.1. Cost-aware consistency management in the cloud

Participants: Houssem-Eddine Chihoub, Shadi Ibrahim, Gabriel Antoniu.

With the emergence of cloud computing, many organizations have moved their data to the cloud in order to provide scalable, reliable and highly available services. To meet ever growing user needs, these services mainly rely on geographically-distributed data replication to guarantee good performance and high availability. However, with replication, consistency comes into question. Service providers in the cloud have the freedom to select the level of consistency according to the access patterns exhibited by the applications. Most optimizations efforts then concentrate on how to provide adequate trade-offs between consistency guarantees and performance. However, as the monetary cost completely relies on the service providers, in [20] we argue that monetary cost should be taken into consideration when evaluating or selecting a consistency level in the cloud. Accordingly, we define a new metric called *consistency-cost efficiency*. Based on this metric, we present a simple, yet efficient economical consistency model, called *Bismar*, that adaptively tunes the consistency level at run-time in order to reduce the monetary cost while simultaneously maintaining a low fraction of stale reads. Experimental evaluations with the Cassandra cloud storage on a Grid'5000 testbed show the validity of the metric and demonstrate the effectiveness of the proposed consistency model.

6.3.2. Analysis of the impact of consistency mangement on energy consumption

Participants: Houssem-Eddine Chihoub, Shadi Ibrahim, Gabriel Antoniu.

Energy consumption within datacenters has grown exponentially in recent years. In the era of Big Data, storage and data-intensive applications are one of the main causes of the high power usage. However, few studies have been dedicated to the analysis of the energy consumption of storage systems. Moreover, the impact of consistency management has never been investigated in spite of its high importance. In this work, we address this particular issue. We investigate the energy consumption of application workloads with different consistency models. Thereafter, we leverage the observations about power and the resource usage with every consistency level in order to provide insight into energy-saving practices. In this context, we introduce adaptive configurations of the storage cluster according to the used consistency level. Our experimental evaluations on Cassandra deployed on Grid'5000 demonstrate the existence of the inevitable tradeoff between consistency and energy saving. Moreover, they show how reconfiguring the storage cluster can lead to energy saving, enhanced performance, and better consistency.

6.3.3. Chameleon: customized consistency by means of application behavior modeling

Participants: Houssem-Eddine Chihoub, Gabriel Antoniu.

Multiple Big Data applications are being deployed worldwide to serve a very large number of clients nowadays. These applications vary in their performance and consistency requirements. Understanding such requirements at the storage system level is not possible. The high level semantics of an application is not exposed at the system level. In this context, the consequences of a stale read are not the same for all types of applications.

In [28] , we focus on managing consistency at the application level rather than at the system level. In order to achieve this goal, we propose an offline modeling approach of the application access behavior that considers its high–level consistency semantics. Furthermore, every application state is automatically associated with a consistency policy. At runtime, we introduce the *Chameleon* approach that leverages the application model to provide a customized consistency specific to that application. Experimental evaluations show the high accuracy of our modeling approach exceeding 96% of correct classification of the application states. Moreover, our experiments conducted on Grid'5000 show that *Chameleon* adapts, for every time period, according to the application behavior and requirements while providing best-effort performance.

6.4. Scalable I/O and Virtualization for Exascale Systems

6.4.1. Damaris/Viz

Participants: Matthieu Dorier, Gabriel Antoniu, Lokman Rahmani.

In the context of the Joint Inria/UIUC/ANL Laboratory for Petascale computing (JLCP), we are developing Damaris, which enables efficient I/O, data analysis and visualization at very large scale from SMP machines. The I/O bottlenecks already present on current petascale systems as well as the amount of data written by HPC applications force to consider new approaches to get insights from running simulations. Trying to bypass the need for storage or drastically reducing the amount of data generated will be of outmost importance for exascale. In-situ visualization has therefore been proposed to run analysis and visualization tasks closer to the simulation, as it runs.

We investigated the limitations of existing in-situ visualization software and proposed Damaris/Viz, a new version of Damaris that fills the gaps of these software by providing in-situ visualization support to Damaris. The use of Damaris/Viz on top of existing visualization packages allows us to:

- Reduce code instrumentation to a minimum in existing simulations,
- Gather the capabilities of several visualization tools to offer adaptability under a unified data management interface,
- Use dedicated cores to hide the run time impact of in-situ visualization and
- Efficiently use memory through a shared-memory-based communication model.

Experiments were conducted on Blue Waters (Cray XK6 at NCSA), Intrepid (BlueGene/P at ANL) and Grid'5000 with representative visualization scenarios for the CM1 [33] atmospheric simulation and the Nek5000 [35] CFD solver. Part of these experiments were carried by NCSA researcher Roberto Sisneros, who gave us important (and very positive) feedbacks on the usability of Damaris at scale (up to 6400 cores on Blue Waters) with real applications. The results of this work were presented as a poster in the PhD forum of IEEE IPDPS 2013 [22], published in a research report [29] and at the IEEE LDAV 2013 conference [23], and a demo of Damaris/Viz was presented at Inria's exhibition booth at the Supercomputing (SC 2013) conference.

This work enlightened the fact that interactive in-situ visualization, although greatly improved by Damaris/Viz, still lakes interactivity. Several meetings were organized with Tom Peterka (ANL) and Roberto Sisneros (NCSA) during the SC conference and during the 10th workshop of the JLPC. We started working on an approach that leverages information theory metrics to automatically find important features of the simulations' data and to reduce the visualization load accordingly.

6.4.2. CALCioM

Participants: Matthieu Dorier, Gabriel Antoniu.

Unmatched computation and storage performance in new HPC systems have led to a plethora of I/O optimizations ranging from application-side collective I/O to network and disk-level request scheduling on the file system side. As we deal with ever larger machines, the interference produced by multiple applications accessing a shared parallel file system in a concurrent manner becomes a major problem. Interference often breaks single-application I/O optimizations, dramatically degrading application I/O performance and, as a result, lowering machine wide efficiency.

Following discussions initiated in 2012 with ANL's Rob Ross and Dries Kimpe, a three month internship of Matthieu Dorier at Argonne National Lab during the summer 2013 led to the design and evaluation of CALCioM (Cross-Application Layer for Coordinated I/O Management), a framework that aims to mitigate I/O interference through the dynamic selection of appropriate scheduling policies. CALCioM allows several applications running on a supercomputer to communicate and coordinate their I/O strategy in order to avoid interfering with one another. Several I/O strategies were evaluated using this framework. Experiments on Argonne's BG/P Surveyor machine and on several clusters of Grid'5000 showed how CALCioM can be used to efficiently and transparently improve the scheduling strategy between several otherwise interfering applications, given specified metrics of machine wide efficiency.

Future work will explore approaches to automatically detect the temporal I/O patterns of simulations in order to further improve the scheduling decisions made by CALCioM.

6.4.3. Scalable metadata management for WAN

Participants: Rohit Saxena, Alexandru Costan, Gabriel Antoniu.

BlobSeer-WAN is a data management service specifically optimized for geographically distributed environments. It is an extension of BlobSeer, a large scale data management service. The metadata is replicated asynchronously for low latency. There is a version manager on each site and vector clocks are used to enable collision detection and resolution under highly concurrent access. It was developed within the framework of Viet-Trung Tran' s PhD thesis, in relation to the FP3C project.

BlobSeer-WAN is used as a storage backend for HGMDS, a multi master metadata server designed for a global distributed file system, developed at University of Tsukuba. Several experiments have been conducted with this setup on the Grid'5000 testbed which have shown scalable metadata performance under geographically distributed environments.

LAGADIC Project-Team

6. New Results

6.1. Visual tracking

6.1.1. 3D model-based tracking

Participants: Antoine Petit, Eric Marchand.

This study focused on the issue of estimating the complete 3D pose of the camera with respect to a potentially textureless object, through model-based tracking. We proposed to robustly combine complementary geometrical and color edge-based features in the minimization process, and to integrate a multiple-hypotheses framework in the geometrical edge-based registration phase [53], [52], [68], [11].

6.1.2. Pose estimation through multi-planes tracking

Participants: Bertrand Delabarre, Eric Marchand.

This study dealt with dense visual tracking robust towards scene perturbations using 3D information to provide a space-time coherency. The proposed method is based on a piecewise-planar scenes visual tracking algorithm which aims at minimizing an error between an observed image and reference templates by estimating the parameters of a rigid 3D transformation taking into acount the relative positions of the planes in the scene. Both the sum of conditional variance and mutual information have been considered[40] [67].

6.1.3. Pose estimation from spherical moments

Participant: François Chaumette.

This study has been realized in collaboration with Omar Tahri from ISR in Coimbra (Portugal) and Youcef Mezouar from Institut Pascal in Clermont-Ferrand. It was devoted to the classical PnP (Perspective-from-N-Points) problem whose goal is to estimate the pose between a camera and a set of known points from the image measurement of these points. We have developed a new method based on invariant properties of the spherical projection model, allowing us to decouple the pose estimation in two steps: the first one provides the translation by minimizing a criterium using an iterative Newton-like method, the second one directly provides the rotation by solving a Procrustes problem [65], [26].

6.1.4. Structure from motion

Participants: Riccardo Spica, Paolo Robuffo Giordano, François Chaumette.

Structure from motion (SfM) is a classical and well-studied problem in computer and robot vision, and many solutions have been proposed to treat it as a recursive filtering/estimation task. However, the issue of *actively* optimizing the transient response of the SfM estimation error has not received a comparable attention. In the work [64], we studied the problem of designing an online active SfM scheme characterized by an error transient response equivalent to that of a reference linear second-order system with desired poles. Indeed, in a nonlinear context, the observability properties of the states under consideration are not (in general) time-invariant but may depend on the current state and on the current inputs applied to the system. It is then possible to simultaneously act on the estimation gains and system inputs (i.e., the camera velocity for SfM) in order to optimize the observation process and impose a desired transient response to the estimation error. The theory developed in [64] has a general validity and can be applied to many different contexts: in [64] it is shown how to tailor the proposed machinery to two concrete SfM problems involving structure estimation for point features and for planar regions from measured image moments.

6.1.5. 3D reconstruction of transparent objects

Participant: Patrick Rives.

This work has been realized in collaboration with Nicolas Alt, Ph.D. student at the "Technische Universität München" (TUM).

Visual geometry reconstruction of unstructured domestic or industrial scenes is an important problem for applications in virtual reality, 3D video or robotics. With the advent of Kinect sensor, accurate and fast methods for 3D reconstruction have been proposed. However, transparent objects cannot be reconstructed with methods that assume a consistent appearance of the observed 3D structure for different viewpoints. We proposed an algorithm that searches the depth map acquired by a depth camera for inconsistency effects caused by transparent objects. Consistent scene parts are filtered out. The result of our method hence complements existing approaches for 3D reconstruction of Lambertian objects [30].

6.1.6. Pseudo-semantic segmentation

Participants: Rafik Sekkal, Marie Babel.

This study has been realized in collaboration with Ferran Marques from Image Processing Group of the Technical University of Catalonia (Barcelona). We designed a video segmentation framework based on contour projections. This 2D+t technique provides a joint hierarchical and multiresolution solution. Results obtained on state-of-the-art benchmarks have demonstrated the ability of our framework to insure the spatio-temporal consistency of the regions along the sequence.

6.1.7. Augmented reality

Participants: Pierre Martin, Eric Marchand.

Using Simultaneous Localization And Mapping (SLAM) methods becomes more and more common in Augmented Reality (AR). To achieve real-time requirement and to cope with scale factor and the lack of absolute positioning issue, we proposed to decouple the localization and the mapping step. This approach has been validated on an Android Smartphone through a collaboration with Orange Labs [46].

Dealing with AR, we have proposed a method named Depth-Assisted Rectification of Patches (DARP), which exploits depth information available in RGB-D consumer devices to improve keypoint matching of perspectively distorted images [44].

6.2. Visual servoing

6.2.1. Photometric moment-based visual servoing

Participants: Manikandan Bakthavatchalam, François Chaumette.

This goal of this work is to use a set of photometric moments as visual features for visual servoing. We first determined the analytical form of the interaction matrix related to these moments. From the results obtained in the past from binary moments, we then selected a set of four features to control four degrees of freedom (dof) with excellent decoupling and stability properties [35]. More recently, thanks to a collaboration with Omar Tahri from ISR Coimbra in Portugal, these results have been extended to the full six dof case.

6.2.2. Visual servoing of humanoid robot

Participant: François Chaumette.

This study has been realized in collaboration with the Pal robotics company located in Barcelona, Spain. It was devoted to the control of the arm of a humanoid robot by visual servoing for manipulation tasks [29].

6.2.3. Visual servoing of cable-driven parallel robot

Participant: François Chaumette.

This study is realized in collaboration with Rémy Ramadour and Jean-Pierre Merlet from Coprin group at Inria Sophia Antipolis. Its goal is to adapt visual servoing techniques for cable-driven parallel robot in order to achieve acurate manipulation tasks. This study is in the scope of the Inria large-scale initiative action PAL (see Section 8.2.6).

6.2.4. Nanomanipulation

Participants: Le Cui, Eric Marchand.

We began a work, within the ANR P2N Nanorobust project (see Section 8.2.1), on the development of micro- and nano-manipulation within SEM (Scanning Electron Microscope). Our goal is to provide visual servoing techniques for positioning and manipulation tasks with a nanometer precision. This year, we focused on the characterisation of the projection model of a SEM along with the approach required for its calibration.

6.3. Visual navigation of mobile robots

6.3.1. New RGB-D sensor design for indoor 3D mapping

Participants: Eduardo Fernandez Moral, Patrick Rives.

A multi-sensor device has been developed for omnidirectional RGB-D (color+depth) image acquisition (see Figure 5 .a). This device allows acquiring such omnidirectional images at high frame rate (30 Hz). This approach has advantages over other alternatives used nowadays in terms of accuracy and real-time spherical image construction of indoor environments, which are of particular interest for mobile robotics. This device has important prospective applications, such as fast 3D-reconstruction or simultaneous localization and mapping (SLAM). A novel calibration method for such device has been developed. It does not require any specific calibration pattern, taking into account the planar structure of the scene to cope with the fact that there is no overlapping between sensors. A method to perform image registration and visual odometry has also been developed. This method relies in the matching of planar primitives that can be efficiently obtained from the depth images. This technique performs considerably faster than previous registration approaches based on ICP.

6.3.2. Long term mapping

Participants: Tawsif Gokhool, Patrick Rives.

This work inscribes in the context of lifelong navigation and map building. The kind of representation that we focus on is made up of a topometric map consisting of a graph of spherical RGB-D views. Thanks to the use of a saliency map built from the photometric and geometric data, we are able to characterize the conditioning of the pose estimation algorithm and to keep as keyframes only a subset of the spherical RGB-D views acquired on the fly. Subsequently, a study on the spread of keyframes was made. The aim was to investigate ways of covering completely and optimally the explored environment in a pose graph representation. Again, over here, the benefits are twofold. Firstly, data acquisition at a throttle of 30 Hz induces many redundant information in the database, which may not necessarily contribute much to the registration phase. Therefore, intelligent selection of keyframes helped in the reduction of data redundancy. Furthermore, as pointed out in the literature, frame to keyframe alignment has the advantage of reducing trajectory drift since the propagation error is diminished as well (see Figure 5 .b)

6.3.3. Semantic mapping

Participants: Romain Drouilly, Patrick Rives.

Semantic mapping aims at building rich cognitive representations of the world in addition to classical topometric maps. A dense labeling has been achieved from high resolution outdoor images using an approach combining Random Forest (RF) and Conditional Random Field (CRF). A second development dealt with the use of semantic information for localization in indoor scenes. For this kind of scenes dense labeling is more difficult due to the large number of potential classes. Therefore algorithms developed for this task rely on a sparse representation of indoor environments called “pbmap”. It consists of a graph whose nodes are the planes present in a given scene. These planes are the only parts of the scene that are labeled. Very high labeling rates of planes has been reached (more than 90%) and it has been shown that these labeled planes could be useful for localization and navigation tasks.

6.3.4. Automous navigation of wheelchairs

Participants: Rafik Sekkal, François Pasteau, Marie Babel.

(a) (b)

Figure 5. a) Omnidirectional RGB-D sensor, b) Top view of dense visual SLAM with fusion of intensity and depth

The goal of this work is to design an autonomous navigation framework of a wheelchair by means of a single camera and visual servoing. We focused on a corridor following task where no prior knowledge of the environment is required. The servoing process matches the non-holonomic constraints of the wheelchair and relies on two visual features, namely the vanishing point location and the orientation of the median line formed by the straight lines related to the bottom of the walls [60]. This overcomes the initialization issue typically raised in the literature. The control scheme has been implemented onto a robotized wheelchair and results show that it can follow a corridor with an accuracy of ± 3 cm [50]. This study is in the scope of the Inria large-scale initiative action PAL (see Section 8.2.6) as well as of the Apash project (see Section 8.1.2).

6.3.5. *Semi-autonomous control of a wheelchair for navigation assistance along corridors*

Participants: Marie Babel, François Pasteau, Alexandre Krupa.

This study concerns a semi-autonomous control approach that we designed for safe wheelchair navigation along corridors. The control relies on the combination of a primary task of wall avoidance performed by a dedicated visual servoing framework and a manual steering task. A smooth transition from manual driving to assisted navigation is obtained thanks to a gradual visual servoing activation method that guarantees the continuity of the control law. Experimental results clearly show the ability of the approach to provide an efficient solution for wall avoiding purposes. This study is in the scope of the Inria large-scale initiative action PAL (see Section 8.2.6) as well as of the Apash project (see Section 8.1.2).

6.3.6. *Target tracking*

Participants: Ivan Markovic, François Chaumette.

This study was realized in the scope of the FP7 Regpot Across project (see Section 8.3.1.2) during the three-month visit of Ivan Markovic, Ph.D. student at the Unviersity of Zagreb. It consisted in developing a pedestrian visual tracking from an omni-directional fish-eye camera and a visual servoing control scheme so that a mobile robot is able to follow the pedestrian. This study has been validated on our Pioneer robot (see Section 5.5).

6.3.7. *Obstacle avoidance*

Participants: Fabien Spindler, François Chaumette.

This study was realized in collaboration with Andrea Cherubini who is now Assistant Prof. at Université de Montpellier. It is concerned with our long term researches about visual navigation from a visual memory without any accurate 3D localization [9]. In order to deal with obstacle avoidance while preserving the visibility in the visual memory, we have proposed a control scheme based on tentacles for fusing the data provided by a pan-tilt camera and a laser range sensor [14]. Recent progresses have been obtained by considering moving obstacles [39].

6.4. Medical robotics

6.4.1. *Needle detection and tracking in 3D ultrasound*

Participants: Pierre Chatelain, Alexandre Krupa.

We developed an algorithm for detecting and tracking a flexible needle in a sequence of 3D ultrasound volumes when it is manually inserted, without any a priori information on the insertion direction. Our approach is based on the combination of a RANSAC algorithm with Kalman filtering in a closed loop fashion and allows real-time tracking of the needle. In addition, a pose-based visual servoing was developed for automatically moving a robotized 3D ultrasound probe in order to keep the needle tip centered in the volume and to align its main axis with the central plane of the volume. This needle detection algorithm and probe automatic guidance were experimentally validated during the insertion of a needle in a gelatin phantom [38].

6.4.2. *Non-rigid target tracking in ultrasound images*

Participants: Marie Babel, Alexandre Krupa.

In order to robustly track the motion of a tumour or cyst during needle insertion, we developed a new approach to track a deformable target within a sequence of 2D ultrasound images. It is based on a dedicated hierarchical grid interpolation algorithm (HGI) that is typically used for real-time video compression purposes. This approach provides a continuous motion representation of the target by using a grid of control points that models both their global displacement and local deformations. The motion of each control point is estimated by a hierarchical and multi-resolution local search method in order to minimize the sum of squared difference of the target pixel intensity between successive images. This new approach was validated from 2D ultrasound images of real human tissues undergoing rigid and non-rigid deformations.

6.4.3. Adaptive arc-based path planning for robot-assisted needle 3D steering using duty-cycling control technique

Participant: Alexandre Krupa.

This study concerned the development of a method for three dimensional steering of a beveled-tip flexible needle that can be used in medical robotics for percutaneous assistance procedures. The proposed solution is the extension of an adaptive arc-based 2D planar approach. It combines the Rapidly-Exploring Random Tree (RRT) algorithm, the duty-cycling needle control technique and stop and turn phases to reorientate the needle in a new working plane each time it is necessary. Simulation results demonstrate the feasibility of this approach to reach a 3D target while avoiding obstacles and its robustness to needle kinematic model errors.

6.4.4. Gait analysis

Participants: Cyril Joly, Patrick Rives.

Clinical evaluation of frailty in the elderly is the first step to decide the degree of assistance they require. Advances in robotics make it possible to turn a standard assistance device into an augmented device that may enrich the existing tests with new sets of daily measured criteria. We designed an augmented 4-wheeled rollator, equipped with a Kinect and odometers, for daily biomechanical gait analysis. It allows to estimate on line legs and feet configurations during the walk. Preliminary results [43] obtained on four healthy persons show that relevant data can be extracted for gait analysis (*e.g.* foot orientation and tibia-foot angle, feet position) during an assisted walk.

This work has been realized in collaboration with Claire Dune from the University of Toulon and in the scope of the Inria large-scale initiative action PAL (see Section 8.2.6).

6.5. Control of single and multiple UAVs

6.5.1. State estimation and flight control of quadrotor UAVs

Participants: Riccardo Spica, Paolo Robuffo Giordano.

Over the last years the robotics community witnessed an increasing interest in the Unmanned Aerial Vehicle (UAV) field. In particular quadrotor UAVs have become more and more widespread in the community as experimental platform for, e.g., testing novel 3D planning, control and estimation schemes in real-world indoor and outdoor conditions. Indeed, in addition to being able to take-off and land vertically, quadrotors can reach high angular accelerations thanks to the relatively long lever arm between opposing motors. This makes them more agile than most standard helicopters or similar rotorcraft UAVs, and thus very suitable to realize complex tasks such as aerial mapping, air pollution monitoring, traffic management, inspection of damaged buildings and dangerous sites, as well as agricultural applications such as pesticide spraying.

Key components for the successful deployment of such systems are (i) a reliable state estimation module able to deal with highly unstructured and/or GPS-denied indoor environments, and (ii) a robust flight control algorithm able to cope with model uncertainties and external disturbances (e.g., adverse atmospheric conditions). The difficulty of these estimation and control problems is also increased by the limited amount of sensing and processing capabilities onboard standard quadrotors: this clearly imposes additional strict requirements on the complexity of the employed algorithms.

In the context of robust flight control of standard quadrotors, the works [31], [32] addressed the theoretical developments and experimental validation of a novel nonlinear adaptive flight controller able to estimate online the UAV dynamic parameters (such as the position of the center of mass when carrying unmodeled payloads), and to compensate for external wind gusts. In parallel, we also developed in [63] a high performance and open-source hardware/software control architecture for flight control of quadrotor UAVs made available to the general public on a open repository. This was achieved by combining state-of-the-art filtering and control techniques with a careful customization and calibration of a commercially available and low-cost quadrotor platform. Finally, still in the context of flight control, the work [58] reported a successful experimental validation of several flight tests for a novel overactuated quadrotor design with tilting propellers behaving as a fully-actuated rigid body in 3D space (thus, able to control its position and orientation in a fully decoupled way).

As for state estimation, the work [41] introduces a novel nonlinear estimation filter meant to obtain a metric measurement of the body-frame linear velocity from optical flow decomposition (thus, visual input) and concurrent fusion of the accelerometer/gyro readings from the onboard IMU. The peculiarity of this filtering technique is the possibility to both explicitly characterize and impose the transient response of the estimation error (thus, the filter performance) by acting on the estimation gains and UAV motion (acceleration). This is in contrast with the consolidated use of EKF schemes which, because of their inherent linearization of the system dynamics, do not typically allow to draw any conclusions about the stability/transient response of the estimation error.

These works were realized in collaboration with the robotics groups at the University of Cassino, Italy, and at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

6.5.2. Collective control of multiple UAVs

Participant: Paolo Robuffo Giordano.

The challenge of coordinating the actions of multiple robots is inspired by the idea that proper coordination of many simple robots can lead to the fulfilment of arbitrarily complex tasks in a robust (to single robot failures) and highly flexible way. Teams of multi-robots can take advantage of their number to perform, for example, complex manipulation and assembly tasks, or to obtain rich spatial awareness by suitably distributing themselves in the environment. Within the scope of robotics, autonomous search and rescue, firefighting, exploration and intervention in dangerous or inaccessible areas are the most promising applications.

In the context of multi-robot (and multi-UAV) coordinated control, *connectivity* of the underlying graph is perhaps the most fundamental requirement in order to allow a group of robots accomplishing common goals by means of *decentralized* solutions. In fact, graph connectivity ensures the needed continuity in the data flow among all the robots in the group which, over time, makes it possible to share and distribute the needed information. In this respect, in [23] a fully decentralized strategy for continuous connectivity maintenance for a group of UAVs has been theoretically developed and experimentally validated on a team of 4 quadrotor UAVs. An extension for allowing an external planner (e.g., a human user) to vary online the minimum degree of connectivity of the group was also proposed in [59]. Finally, [48] dealt with the issue of coupling the purely reactive strategy for connectivity maintenance with an autonomous exploration algorithm in a cluttered 3D environment (still experimentally tested on a team of quadrotor UAVs). The complete software architecture developed for performing these and similar multi-UAV experiments was also published in [42].

These works were realized in collaboration with the robotics group at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

MIMETIC Project-Team

6. New Results

6.1. Biomechanics and Motion Analysis

6.1.1. Modeling gesture in sports: tennis serve

Participants: Nicolas Bideau, Guillaume Nicolas, Benoit Bideau, Richard Kulpa.

In the midst of the INSEP project and the PhD of Caroline Martin, the tennis serve has been studied with biomechanical analyses. To this end, we have done kinematic and dynamic analyses based on motion capture, force plate and electromyographic systems. They provided information on how the gesture is performed and how it is related to injuries. Moreover, these analyses have been done on several level of players including top-level ones. A comparison of the kinematic and dynamic data can then be done. Our objective is to use these data in virtual reality to study the interaction between a tennis server and a receiver. We are creating a tool that displays a virtual server in front of a real receiver. The control of the virtual server is then done based on these biomechanical data. The objective is to analyze the reaction of the receiver depending on the movement of the server and its level of expertise.

6.1.2. Motion modeling in clinical applications

Participant: Armel Crétual.

We have developed a new index of gait quantification based on muscular activity called KeR-EGI. After having proved that this index is consistent and complementary with kinematics-based indices, we have shown that it is reproducible in patients even when their impairement level is high. This index is now used in clinical routine in adults. It will be also used in pediatrics in the next few months.

In orthopedics, we have proposed a novel method to quantify shoulder's global mobility called SCSV. It is based on the reachable volume in the whole configuration space of the shoulder, i.e. a 3-dimensional angular space. Clinical evaluations of shoulder's range of motion are quite always based on the analysis of only one axis, and the most usual refers to maximal external rotation from rest posture (ER1). Considering several mono-axial amplitudes, we have shown that ER1 is actually the worst choice to estimate global mobility. Instead of the ER1 procedure, we proposed to use the sum of 3 mono-axial amplitudes: external/internal amplitude at 90° lateral elevation, abduction and flexion/extension.

As shoulder is actually a complex of three articulations (gleno-humeral, scapulo-thoracic and sterno-calvicular), we have evaluated the contribution of each of them on global mobility. This has been done through a cadaveric study where we measured SCSV in any possible blocking conditions of these three articulations (from 0 to 3).

6.2. VR and Ergonomics

Participants: Charles Pontonnier [contact], Georges Dumont, Franck Multon, Pierre Plantard.

The use of virtual reality tools for ergonomics applications is a very important challenge in order to genezalize the use of such devices for the design of workstations.

First, the development of motion analysis tools is mandatory in order to provide additional information to the ergonomists and help them to analyse the work environment. Particularly, an analysis of the muscle forces involved in the motion generation is a very important information with regard to the ergonomics of a task. Several methods can lead to an estimation of these muscle forces. In a study we developed, we tried to assess the level of confidence for results obtained with an inverse dynamics method from real captured work tasks. The chosen tasks were meat cutting tasks, well known to be highly correlated to musculoskeletal troubles appearance in the slaughter industry.

The experimental protocol consisted in recording three main data during meat cutting tasks, and analysing their variation when some of the workstation design parameters were changing.

Then the motion was replayed in the AnyBody modeling system (AnyBody, Aalborg, Denmark) in order to obtain muscle forces generated during the motion. A trend comparison has been done, comparing recorded and computed muscle activations. Results showed that most of the computed activations were qualitatively close from the recorded ones (similar shapes and peaks), but quantitative comparison leaded to major differences between recorded and computed activations (the trend followed by the recorded activations in regard of a workstation design parameter, such as the table height, is not obtained with the computed activations) [15]. We currently explore those results to see if the fact that co-contraction of single joints muscles is badly estimated by classical inverse dynamics method can be a reason of this issue. We also work on the co-contration simulation in order to improve the results.

This work has been done in collaboration with the Center for Sensory-motor Interaction (SMI, Aalborg University, Aalborg, Denmark), particularly Mark de Zee (Associate Professor) and Pascal Madeleine (Professor).

Furthermore, the fidelity of the VR simulator has to be evaluated. For example, a simulator for assembly task has been evaluated in comparing different types of interaction : real, virtual and virtual + force feedback. Objective and subjective metrics of discomfort led to highlight the influence of the environment on motor control and sensory feedback, changing more or less deeply the way the task is performed. The results particularly showed a distorsion between the user's subjective rating of discomfort and the objective value associated to the postures they reached during the task execution. Nevertheless, scores obtained in real and virtual environments for objective and subjective indicators of discomfort were highly correlated [17], [16]. It indicates that despite the differences, the gap between real and virtual environments can be fullfiled. This work has been done within the frame of the european project FP7 VISIONAIR.

At last we proposed in collaboration with Thierry Duval (Hybrid team, Rennes) a new architecture for information sharing and bridging in collaborative virtual environements in application to ergonomics studies. This work has been awarded with a best paper award at The 4th IEEE conference on Cognitive Infocommunications (CogInfoCom 2013) [28].

6.3. Motion Sensing and analysis

Participant: Franck Multon [contact].

Sensing human activity is a very active field of research, with a wide range of applications ranging from entertainment and serious games to personal ambient living assistance, including rehabilitation. MimeTIC aims at proposing original methods to process raw motion capture data in order to compute relevant information according to the application.

In rehabilitation, we have collaborated with University of Montreal, Saint-Justine Hospital which main activity is rehabilitation of children with pathologies of the pyramidal control system. In this domain, defining metrics and relevant measurement to diagnose pathologies and to monitor patients during treatment is a key point. In gait, most of the previous works focus on gait spatio-temporal parameters (such as step length, frequency, stride duration, global speed) which could be measured with two main families of systems: 1) one-point measurement with a force plate, one accelerometer or dedicated devices (such as a Gait Ride), or 2) multi-point measurement systems with motion sensors or markers placed over the patient's skin. The former provides the clinician with compact but incomplete knowledge whereas the latter provides him with numerous data which are sometimes difficult to analyze and to get (specific technical skills are required). The first step to any type of analysis is to detect the main gait events, such as foot strikes and toe offs. In treadmill walking, widely used in rehabilitation as it enables the clinician to analyze numerous gait cycles in a limited place with a controlled speed, automatically detecting such gait events requires complex devices with specific technical skills (such as calibration and post-processing with motion capture systems).

Recent papers have demonstrated that low-cost and easy-to-use depth cameras (such as a Kinect from Microsoft) look promising for serious applications requiring motion capture. However there exist some confusion between the feet and the ground at foot strike and foot off leading to bad estimation of the gait cycle events. We have proposed an alternative approach that consists in using the strong correlation between knee and foot trajectories to deduce foot strikes thanks to knee movements. The extremes of the distance between the two knees along the longitudinal axis provides us with very accurate gait events detection compared to previous works.

A second contribution consisted in defining a global gait asymmetry index according to depth images provided by a Kinect. In previous works this index relied on computing ratio between joint angles. With a Kinect, joint angles may be very noisy that could affect the asymmetry index. We have introduced a new index which is directly deduced from depth images without any joint angle estimation nor skeleton fitting. The method consists in building a model of the gait cycle of the patient by averaging depth images recorded along several cycles. As a consequence the noise within the instantaneous depth images is filtered leading to accurate surfaces of the patient gait (leading to a 3D+time data structure). The main vertical axis of the surface is used to define a symmetry plane. Consequently surfaces of the right part of the body can be symmetrized to be compared to the left part at compatible times in the gait cycle (such as a right foot strike is symmetrized to be compared to a left foot strike). The comparison between the two surfaces leads to a promising asymmetry index. The results (see Figure 4) demonstrate that this method is able to significantly distinguish asymmetrical gaits obtained by adding a 5cm sole under one of the feet of healthy subjects. Ongoing works consist in comparing this index to previously published ones which were based on accurate motion capture data. It will also be applied to unimpaired gaits of pathological subjects.

Figure 4. Longitudinal (DAI) and lateral (LAI) Asymmetry indexes computed thanks to depth images for normal gait and two artificial modification (adding a 5cm sole below the left or the right foot). The asymmetry index is computed all along the gait cycle and was able to statistically distinguish asymmetrical gaits.

6.4. VR and Sports

Participants: Richard Kulpa [contact], Benoit Bideau, Franck Multon.

Previous works in MimeTIC have shown the advantage of using VR to design and carry-out experiments on perception-action coupling in sports, especially for duels between two opponents. However the impact of using various technical solutions to carry-out this type of experiment in sports is not clear. Indeed immersion is performed by using interfaces to capture the motion/intention of the user and to deliver various multi-sensory feedbacks. These interfaces may affect the perception-action loop so that results obtained in VR cannot be systematically transferred to real practice.

Most of the applications in VR provide the user with visual feedbacks in which the avatar of the user can be more or less simplified (sometimes limited to a hand or the tools he his carrying). In first person view in caves the user generally does not need accurate avatars as he can perceive his real body but some authors have shown that the perception of distances is generally modified. Some authors have also demonstrated that first-person view was less efficient that third person view with avatars when performing accurate tasks such as reaching objects in constrained environments. We proposed an experiment to evaluate which type of feedback was the most appropriate one for complex precision tasks, such as basketball free-throw. In basketball free-throw the user has to throw a ball into a small basket placed at over 4.5m far from him. Thus perception of distance is actually a key point in such a task. Beginners and experts carried-out a first experiment in real in order to measure their motion and performance in real situation. Then beginners were asked to perform free throws with a real ball in hands, but in three conditions in a Cave (Immersia Room, Rennes): 1) first-person view (see Figure 5), 2) third-person view with the visual feedback of the ball's position, and 3) third-person view the virtual ball and additional rings modeling the perfect trajectory for the ball to get in the basket. Results show that significant difference exists in ball speed between first-person view condition compared to real condition whereas no difference exist in third-person view conditions. If we focus on successful throws only, ball speed in the last condition 3) was very similar to real condition whereas all the other VR conditions (1) and 2)) lead to significant differences compared to real situation. In all VR conditions the height of ball release was significantly higher in VR compared to real situation. These results show that VR conditions lead to adaptations in the way people perform such a precision task, especially for ball speed and height of ball release. However this difference is significantly higher with first person view and tends to zero in condition 3). Future works will tend to evaluate new conditions with avatars and complementary points of view (such as lateral and frontal views together as suggested by some authors). It will also be important to more clearly understand the problem of perception of distances in such an environment. This work has been performed in cooperation with University of Brassov in Romania.

Another key feedback is the external forces associated with the task. In most sports applications such forces are strongly linked to performance. However delivering these forces in virtual environments is still a challenge as it required haptic devices that could affect the way the users perform the task (with a different grip compared to real situation and limitations in dynamic response of the device). Pseudohaptics has been introduced in the early 2000. It consists in using visual feedbacks to make people perceive the forces linked to a task. However this approach has not been tested for whole-body interaction. In collaboration with Hybrid team in Inria Rennes, we studied how the visual animation of a self-avatar could be artificially modified in real-time in order to generate different haptic perceptions. In our experimental setup participants could watch their self-avatar in a virtual environment in mirror mode. They could map their gestures on the self-animated avatar in real-time using a Kinect. The experimental task consisted in a weight lifting with virtual dumbbells that participants could manipulate by means of a tangible stick. We introduce three kinds of modification of the visual animation of the self-avatar: 1) an amplification (or reduction) of the user motion (change in C/D ratio), 2) a change in the dynamic profile of the motion (temporal animation), or 3) a change in the posture of the avatar (angle of inclination). An example is depicted in Figure 6 . Thus, to simulate the lifting of a "heavy" dumbbell, the avatar animation was distorted in real-time using: an amplification of the user motion, a slower dynamics, and a larger angle of inclination of the avatar. We evaluated the potential of each technique using an ordering task with four different virtual weights. Our results show that the ordering task could be well achieved with every technique. The C/D ratio-based technique was found the most efficient. But participants globally appreciated all the different visual effects, and best results could be observed in the combination configuration. Our results pave the way to the exploitation of such novel techniques in various VR applications such as for sport training, exercise games, or industrial training scenarios in single or collaborative mode.

Figure 5. First-person view condition in the basket free-throw performed in a cave (Immersia Room, France).

Figure 6. Weight discrimination task: the animation of the avatar showed a lifting effort according to the weight of the virtual dumbbell and the user has to rank the conditions from the lightest to the heaviest mass.

6.5. Autonomous Virtual Humans

6.5.1. Space and Time Constrained Task Scheduling for Crowd Simulation

Participants: Carl-Johan Jorgensen, Fabrice Lamarche [contact].

Crowd distribution in cities highly depends on how people schedule their daily activities. When performing an intended activity, people decisions and behavior mainly consist in scheduling tasks that compose this activity, planning paths between locations where these tasks should be performed, navigating along the planned paths and performing the scheduled tasks.

We proposed a task scheduling model aims at selecting where, when and in which order several tasks, representing an intended activity, should be performed. The proposed model handles spatial and temporal constraints relating to the environment and to the agent itself. Personal preferences, characterizing the agent, are also taken into account. Produced task schedules are optimized on the long term and exhibit adequate choices of locations and times with respect to the agent intended activity and its environment. Once computed, these task schedules are relaxed and used to drive a microscopic crowd simulation in which observable flows of pedestrians emerge from the scheduled individual activities. Such simulations are easy to produce and do not require the use of a complex decisional model. In terms of validation, we conducted an experiment that shows that our algorithm produces task schedules which are representative of humans' ones.

This work is part of the iSpace&Time project in which virtual cities are populated with virtual pedestrians and vehicles.

6.5.2. Long term planning and opportunism

Participants: Philippe Rannou, Fabrice Lamarche [contact].

Autonomous virtual characters evolve in dynamic virtual environments in which changes may be unpredictable. One main problem when dealing with long term action planning in dynamic environment is that an agent should be able to behave properly and adapt its behavior to perceived changes while still fulfilling its goals.

We propose a system that combines long term action planning with failure anticipation and opportunism. The system is based on a modified version of an HTN planning algorithm. It generates plans enriched with information that enable a monitor to detect relevant changes of the environment. Once such changes are detected, a plan adaptation is triggered. Such adaptations include modifying the plan to react to a predicted failure and more importantly to exploit opportunities offered by the environment.

This system has been extended to better take into account the relationship between action planning and the environment. It is now combined with our space and time constrained tasks scheduling system (Cf. 6.5) to optimize the choice of locations where actions should be performed.

6.6. Interactive Virtual Cinematography

Participants: Marc Christie [contact], Christophe Lino, Cunka Sanokho.

The domain of Virtual Cinematography explores the operationalization of rules and conventions pertaining to camera placement, light placement and staging in virtual environments. Within the context of the ANR CHROME project, we have tackled the problem of portraying events in complex crowd simulations using steering behaviors. The system we proposed relies on Reynolds' model of steering behaviors to control and locally coordinate a collection of camera agents similar to a group of reporters. In our approach, cameras are either in a scouting mode, searching for relevant events to convey, or in a tracking mode following one or more unfolding events. The key benefit, in addition to the simplicity of the steering rules, holds in the capacity of the system to adapt to the evolving complexity of crowd simulations by self-organizing the cameras to track most of the events. The results have been presented as the Motion in Games conference [21].

We have also created a table-top interactive application to offer collaborative and high-level control on multi-dimensional and temporal data. This has been applied to the collaborative control of cinematographic parameters in a virtual movie, using our cinematographic engine [26].

In the ANR project Cinecitta, we have proposed means to evaluate the sens of balance in synthetic shots. Balance represents the equilibrium of visual weights in the screen, *i.e.* equilibrium of the visual interests one perceives. Balance is a key criteria in the aesthetics of a shot, and only a few approaches have seriously tackled this issue. In our approach, we rely on a dataset of well-balanced shot extracted from real movies to construct a balance feature space. The balance feature space is then used to estimate the sense of balance in new synthetic shot. We have furthermore extended the approach by automatically recomputing viewpoints to improve balance. A journal paper is under submission in Computer Graphics Forum.

6.7. Interactive Stroytelling

Participants: Marc Christie [contact], Hui-Yin Wu.

In 2013, within the Inria Associate Team FORMOSA (see 8.3.1.1), we have proposed a framework for the creation of parametrable and personalized stories in interactive storytelling. In any kind of storytelling, the success of the story relies both on the intricate plot design and control of the author as well as the emotional feedback of the user. With the assistance of computing algorithms combined with the maturing understanding of narrative structures, it is possible for interactive stories to create a more personalized, engaging, and well-controlled narrative content to users than traditional linear narrative. And with the emergence of new storytelling technologies, critical issues concern the creation of such complex narratives in virtual 3D environments, and the coherent simulation of these interactive narratives.

In the framework we proposed, the author can specify characteristics on the story structure and fragments (pieces of story) in order to generate variations of interactive stories. The characteristics we consider are genre, story complexity, and Chatman's modes of plot (eg a good hero fails). The story generation model we devised combines a branching story structure with a three-step graph traversal algorithm that filters and recombines story fragments from the characteristics, generating a high-level interactive script that satisfies all authorial constraints, and provides sufficient abstraction from the technical implementation. The script is then simulated in a real-time storytelling system, featuring autonomous characters and automatic camera control. The work has been presented as a short paper in the CASA conference [30].

We then extended this approach to handle temporal aspects of discourse in stories (i.e. how to temporally rearrange fragments of a story while maintaining consistency and logic whatever the user's choices). By rewriting our graph traversal algorithm (which filters inconsistent branches, and propagates constraints along the branches), and performing the graph traversal on each choice selected by the user, we enable the simulation of consistent temporal variations in stories. This typically allows the creation of flashbacks, flashforwards, parallel and embedded stories. Early results have been presented as a poster as Motion in Games 2013 [21].

6.8. Haptic Cinematography

Participant: Marc Christie [contact].

In 2013, we have demonstrated an approach to Haptic Cinematography in very selective events (2013 CHI conference [40], 2013 Siggraph Emerging Technology [41], 2013 UIST conference [39]). This is joint work with members of the Hybrid team (Anatole Lécuyer, Fabien Danieau) and members of the Technicolor Company (Philippe Guillotel, Nicolas Mollet, Julien Fleureau). Haptic cinematography consists in enhancing our audio-visual experience of movies by adding haptic effects related to the semantics of camera motions. Camera motions in movies, which are typically non-diegetic elements in a narrative, tend to enhance user experience both visually and emotionnaly. The questions we address here are (i) whether the coupling between camera motions and haptic motions improve this experiences and (ii) what are the rules and recommendations for coupling these motions. Results, that we published in the IEEE Multimedia journal [8], demonstrate that (i) the coupling is effective when precisely synchonized, (ii) the direction of motions between the camera and the haptic motions do not need to be correlated, and (iii) haptic metaphores can easily be perceived by the spectators. This opens great perspectives as to how haptic devices can enhance audio-visual contents in more sublte ways than straightforward mappings between diegatic elements and haptic motions.

6.9. Biomechanics for avatar animation

Participants: Julien Pettré [contact], Charles Pontonnier, Georges Dumont, Franck Multon, Ana Lucia Cruz Ruiz, Steve Tonneau.

Bio-inpired controllers and planners are compelling for avatar animation. We are currently engaging several works on the subject within the frame of the ENTRACTE project 8.1.5 .

Ana-Lucia Cruz-Ruiz has been recruited as a PhD student since november 2013 to begin to work on musculoskeletal-based methods for avatar animation. More precisely, the goal of this thesis is to define and evaluate a modular and multiscale whole-body musculoskeletal model usable to analyze and human movement and synthetize realistiv avatar animations. The specificity of the subject is hidden in the words "modular" and "multiscale". "Modular" says that the model has to be easily tunable to be modified in accordance with the investigated motor control theories (uncontrolled manifold, motor synergies,...). "Multiscale" means that the model has to exhibit multiple levels of details cohabiting at the same time, depending on the region of interest investigated. At last, the model have to be easily scalable, in order to be applied to different morphologies. Moreover, she currently explores musculoskeletal-based simplified joint behaviors to improve torque-based dynamics applications.

We also address the problem of planning human motion in constrained environment. In previous approach, planning human motion is performed based on robotics planning algorithms the objective of which is to avoid obstacles. In our approach, we suggest that creating contacts with the obstacles of the environment is actually a mean to perform a motion tasks. We thus model human motion as a sequence of contacts between humans and obstacles. A contact planner is being developed, and results being prepared for publication.

6.10. Crowds

Participants: Julien Pettré [contact], Anne-Hélène Olivier, Julien Bruneau, Jonathan Perrinet, Kevin Jordao, David Wolinski.

6.10.1. Analysis of Locomotion Trajectories during Collision Avoidance

The experimental observation of physical interactions between real walkers is for us a great source of inspiration for the design of realistic microscopic models of crowd simulation. This year, we have continued analysing locomotion trajectories of real walkers during collision avoidance tasks. Analysis focused on individual strategies and role set to solve such a reciprocal interaction. Our analysis revealed that walkers combine re-orientation and speed adaptations to avoid collisions, but more importantly, that the strategies, as well as the global amount of adaptations is dependent on the role each one has in the avoidance (e.g., passing first, giving way). Our results are reported in [13]. In addition, we inspected the role of psychologic factors on the metrics of interactions [27].

6.10.2. Evaluation of Locomotion Trajectories performed in Virtual Reality

Virtual Reality rooms are physically limited in space, and prevent users virtually walking by really walking in larger virtual spaces: a locomotion interface is employed to overcome this issue. The interface is composed of a peripheral device, such as joystick, as well as of a software component which transform users' actions on the peripheral device into a virtual locomotion. In this work, we wondered if users where performing similar trajectories in virtuo than in vivo: such question is important when aiming at using VR form motion analysis purpose. We evaluated the bias introduced by several couples of devices and software components during the execution of goal directed locomotion tasks. As reported in [7], impressive similarities on the formed trajectories even when the device control motions are radically different in comparison with walking motions.

6.10.3. Virtual Populations for large-scale digital environments and Cultural Heritage Applications

We are developing techniques dedicated to the animation of large virtual populations at very low computational cost based on the crowd patches techniques. Crowd patches can be described as 3D animated textures that small

groups animations. They are composed in space to form large population. This year, we coupled the crowd patches approach with mutable shape models: such association enable users cdesigning patches composition in an interactive manner, as introduced in [22]. We applied those techniques to design populations of some old Malaysian trading ports [25].

6.10.4. Macroscopic derivations of microscopic simulation models

Crowd phenomenon exhibit macroscopic structures which derive from the combination of local interactions between individuals. Together with the IMT in Toulouse in the frame of the ANR-Pedigree project (term. 2012), the microscopic models developed in our team has been derived into macroscopic models to demonstrate their ability to provoke the mergence of some typical macroscopic structures [35], [36].

MYRIADS Project-Team

6. New Results

6.1. Dependable Cloud Computing

Participants: Roberto-Gioacchino Cascella, Stefania Costache, Florian Dudouet, Eugen Feller, Filippo Gaudenzi, Yvon Jégou, Ancuta Iordache, David Margery, Christine Morin, Anne-Cécile Orgerie, Guillaume Pierre, Nikos Parlavantzas, Yann Radenac, Matthieu Simonin, Cédric Tedeschi.

6.1.1. Multi-data Center and Multi-cloud

6.1.1.1. Deployment of distributed applications in a multi-provider environment

Participants: Roberto-Gioacchino Cascella, Stefania Costache, Florian Dudouet, Piyush Harsh, Filippo Gaudenzi, Yvon Jégou, Christine Morin.

The move of users and organizations to Cloud computing will become possible when they will be able to exploit their own applications, applications and services provided by cloud providers as well as applications from third party providers in a trustful way on different cloud infrastructures. In the framework of the Contrail European project [39] [50], we have designed and implemented the Virtual Execution Platform (VEP) service in charge of managing the whole life cycle of OVF distributed applications under Service Level Agreement rules on different infrastructure providers [51]. In 2013, we designed the CIMI inspired REST-API for VEP 2.0 with support for Constrained Execution Environment (CEE), advance reservation and scheduling service, and support for SLAs [55], [54] [56]. We integrated support for delegated certificates and developed test scripts to integrate the Virtual Infrastructure Network (VIN) service. VEP 1.1 was slightly modified to integrate the usage control (Policy Enforcement Point (PEP)) solution developed by CNR. The CEE management interface was developed during 2013 and is available through the graphical API as well as through the RESTful API.

6.1.1.2. Towards a distributed cloud inside the backbone

Participants: Anne-Cécile Orgerie, Cédric Tedeschi.

The DISCOVERY proposal currently in phase of construction and lead by Adrien Lèbre from ASCOLA team, and currently on leave at Inria aims at designing a distributed cloud, leveraging the resources we can find in the network's backbone. [3]

In this context, and in collaboration with ASCOLA and ASAP teams, we started the design of an overlay network whose purpose is to be able, with a limited cost, to locate geographically-close nodes from any point of the network. The basis for this overlay is described as part of a recent research report [44].

6.1.1.3. Multi-cloud application deployment in ConPaaS

Participants: Guillaume Pierre, Yann Radenac.

We extended ConPaaS to support application deployment over multiple clouds. There are two main reasons for this: first, it is a necessary mechanism to allow application migration from one cloud to another, without any service interruption. Second, for some applications it may be useful to execute over multiple clouds on a permanent basis, for reliability reasons for example. The main challenges to address were ensuring full network connectivity between resources acquired in multiple clouds. We addressed these issues by integrating the IPOP virtual network in ConPaaS. Second, we designed protocols to ensure application and data migration without any service interruption during the migration.

6.1.2. Scalability of Snooze Self-healing Cloud Management System

Participants: Eugen Feller, Yvon Jégou, David Margery, Christine Morin, Anne-Cécile Orgerie, Matthieu Simonin.

[3] The DISCOVERY website: http://beyondtheclouds.github.io

We evaluated the scalability and resilience of Snooze IaaS management system [26]. Unlike existing systems, for scalability, ease of configuration, and high availability, Snooze is based on a self-organizing and self-healing hierarchical architecture of system services [36], [27], [27]. In Snooze hierarchy, each compute server is managed by a local controller that interacts with one of the group managers to which it is dynamically assigned and the set of group managers is coordinated by a group leader elected among them. We performed an extensive scalability study of Snooze across over 500 servers of the Grid'5000 experimentation testbed. We evaluated the Snooze self-organizing and self-healing hierarchy with thousands of system services. The results show that the resource consumption of the Snooze system services is bounded both during the hierarchy construction and system operation. We also show that Snooze prototype implementation is robust enough to manage thousands of servers and hundreds of VMs. Moreover, its autonomic behavior allows to achieve high availability in the presence of a large number of simultaneous system services failures. Indeed, as long as at least two group managers remain operational the system remains alive. We also demonstrated the application deployment scalability across hundreds VMs on the example of a Hadoop MapReduce application. We participated in the Scale Challenge organized in the framework of the ACM/IEEE CC-Grid 2013 conference [26] and won the second prize.

6.1.3. Application Performance Modeling in Heterogeneous Cloud Environments

Participants: Ancuta Iordache, Guillaume Pierre.

Heterogeneous cloud platforms offer many possibilities for applications for make fine-grained choice over the types of resources they execute on. This opens for example opportunities for fine-grain control of the tradeoff between expensive resources likely to deliver high levels of performance, and slower resources likely to cost less. We designed a methodology for automatically exploring this performance vs. cost tradeoff when an arbitrary application is submitted to the platform. Thereafter, the system can automatically select the set of resources which is likely to implement the tradeoff specified by the user. A publication on this topic is currently in preparation.

6.1.4. Flexible SLA & SLO Management

Participants: Stefania Costache, Christine Morin, Nikos Parlavantzas.

Merkat is a market-based, SLO-driven, PaaS system for private clouds. Merkat dynamically shares resources between competing applications to ensure a fair resource utilization in terms of application priority and actual resource needs. Resources are allocated through a proportional-share auction while autonomous controllers apply elasticity rules to scale application demand according to resource availability and user priority. Merkat provides users the flexibility to adapt controllers to their application types, and it can support diverse application types and performance goals. Merkat is implemented in Python and uses OpenNebula for virtual machine operations.

We evaluated Merkat in simulation and we analyzed the behavior of the system for multiple user types [23]. Furthermore, we deployed Merkat on Grid'5000 and EDF's tested and tested it with applications representative to EDF [22]. Results showed that: (i) the system provides flexible support for different application types (static and malleable) and different SLOs (deadline and performance); (ii) the system provides good user satisfaction achieving acceptable performance degradation, compared to existing centralized solutions. Furthermore, we extended Merkat to manage different clusters and run MPI applications on them. We also submitted a survey on evolution of resource management systems for shared virtualized computing infrastructures to an international journal. This work was carried out in the framework of Stefania Costache's PhD thesis [11].

6.2. Heterogeneous Resource Management

Participants: Eliya Buyukkaya, Djawida Dib, Eugen Feller, Tran Ngoc Minh, Christine Morin, Nikos Parlavantzas, Guillaume Pierre.

6.2.1. Cross-resource scheduling in heterogeneous cloud environments

Participants: Eliya Buyukkaya, Tran Ngoc Minh, Guillaume Pierre.

Allocating resources to applications in a heterogeneous cloud environment is harder than in a homogeneous environment. In a heterogeneous cloud some rare resources are more precious than others, and should be treated carefully to maximize their utilization. Similarly, applications may request groups of resources that exhibit certain inter-resource properties such as the available bandwidth between the assigned resources. We are currently investigating scheduling algorithms for handling such scenarios.

6.2.2. Maximizing private cloud provider profit in cloud bursting scenarios

Participants: Christine Morin, Djawida Dib, Nikos Parlavantzas.

Current PaaS offerings either provide no support for SLA guarantees or provide limited support targeting a restricted set of application types. To overcome this limitation, we are developing an open, SLA-driven PaaS system, called Meryn, that aims at providing SLA guarantees to diverse application types while maximizing the PaaS provider profit. Meryn supports cloud bursting and applies a decentralized protocol for selecting cloud resources, trying to minimize the cost of running applications without affecting their agreed quality properties. We have performed a preliminary evaluation of Meryn [24] and worked on optimising the system and performing further experiments on the Grid5000 testbed. This work is part of Djawida Dib's PhD thesis.

6.2.3. Data life-cycle management in clouds

Participants: Eugen Feller, Christine Morin.

Infrastructure as a Service (IaaS) clouds provide a flexible environment where users can choose and control various aspects of the machines of interest. However, the flexibility of IaaS clouds presents unique challenges for storage and data management in these environments. Users use manual and/or ad-hoc methods to manage storage and data in these environments. FRIEDA is a Flexible Robust Intelligent Elastic Data Management framework that employs a range of data management strategies approaches in elastic environments. In the context of the DALHIS associate team [4], we evaluated the importance of this framework on multiple cloud testbeds. Our evaluation showed that storage planning needs to be performed in coordination with compute planning and the specific configuration of virtual machine had a strong impact on the application (e.g., some applications performed better on small instances than large instances) [40].

6.3. Energy-efficient Resource Infrastructures

Participants: Alexandra Carpen-Amarie, Bogdan Florin Cornea, Ismael Cuadrado Cordero, Djawida Dib, Eugen Feller, Yunbo Li, Christine Morin, Anne-Cécile Orgerie, Guillaume Pierre.

6.3.1. Energy-efficient IaaS clouds

Participants: Alexandra Carpen-Amarie, Christine Morin, Anne-Cécile Orgerie.

Energy consumption has always been a major concern in the design and cost of data centers. The wide adoption of virtualization and cloud computing has added another layer of complexity to enabling an energy-efficient use of computing power in large-scale settings. Among the many aspects that influence the energy consumption of a cloud system, the hardware-component level is one of the most intensively studied. However, higher-level factors such as virtual machine properties, their placement policies or application workloads may play an essential role in defining the power consumption profile of a given cloud system. In this work, we explored the energy consumption patterns of Infrastructure-as-a-Service (IaaS) cloud environments under various synthetic and real application workloads. For each scenario, we investigated the power overhead triggered by different types of virtual machines, the impact of the virtual cluster size on the energy-efficiency of the hosting infrastructure and the tradeoff between performance and energy consumption of MapReduce virtual clusters through typical cloud applications [21].

6.3.2. Energy-aware IaaS-PaaS co-design

Participants: Alexandra Carpen-Amarie, Djawida Dib, Guillaume Pierre, Anne-Cécile Orgerie.

[4] http://project.inria.fr/dalhis

The wide adoption of the cloud computing paradigm plays a crucial role in the ever-increasing demand for energy-efficient data centers. Driven by this requirement, cloud providers resort to a variety of techniques to improve energy usage at each level of the cloud computing stack. However, prior studies mostly consider resource-level energy optimizations in IaaS clouds, overlooking the workload-related information locked at higher levels, such as PaaS clouds. We argue that cross-layer cooperation in clouds is a key to achieving an optimized resource management, both performance and energy-wise. To this end, we claim there is a need for a cooperation API between IaaS and PaaS clouds, enabling each layer to share specific information and to trigger correlated decisions. We identified the drawbacks raised by such co-design objectives and discuss opportunities for energy usage optimizations, and plan to start the research to address these issues in 2014.

6.3.3. *Performance and energy-efficiency evaluation of Hadoop deployment models*

Participants: Eugen Feller, Christine Morin.

The exponential growth of scientific and business data has resulted in the evolution of the cloud computing and the MapReduce parallel programming model. Cloud computing emphasizes increased utilization and power savings through consolidation while MapReduce enables large scale data analysis. The Hadoop framework is the most popular open source software implementing the MapReduce model. In our work, we evaluated Hadoop performance in two modes – the traditional model of collocated data and compute services and separated mode where the services are deployed on separate services. The separation of data and compute services provides more flexibility in environments where data locality might not have a considerable impact such as virtualized environments and clusters with advanced networks. In this work, we also conducted an energy efficiency evaluation of Hadoop on physical and virtual clusters in different configurations. The experiments were performed on the Grid'5000 experimentation testbed. To enable virtual machine management, the Snooze cloud stack developed by the Myriads project-team was used. Our extensive evaluation shows that: (1) performance on physical clusters is significantly better than on virtual clusters; (2) performance degradation due to separation of the services depends on the data to compute ratio; (3) application completion progress correlates with the power consumption and power consumption is heavily application specific [28].

6.3.4. *Energy consumption models and predictions for large-scale systems*

Participant: Christine Morin.

Responsible, efficient and well-planned power consumption is becoming a necessity for monetary returns and scalability of computing infrastructures. While there is a variety of sources from which power data can be obtained, analyzing this data is an intrinsically hard task. In our work, we described a generic approach to analyze large power consumption datasets collected from computing infrastructures. As a first step, we proposed a data analysis pipeline that can handle the large-scale collection of energy consumption logs, apply sophisticated modeling to enable accurate prediction, and evaluate the efficiency of the analysis approach. We presented the analysis of a power consumption data set collected over a 6-month period from two clusters of the Grid'5000 experimentation platform used in production. We used Hadoop with Pig to handle the large volume of data. Our data processing generated a summary of the data that provides basic statistical aggregations, over different time scales. The aggregate data was then analyzed as a time series using sophisticated modeling methods with R statistical software. We exploited time series to detect outliers and derive hourly and daily power consumption predictive models. We demonstrated the accuracy of the predictive models and the efficiency of the data processing performed on a 55-node cluster at NERSC [34]. Energy models from such large dataset can help in understanding the evolution of consumption patterns, predicting future energy trends, and providing basis for generalizing the energy models to similar large-scale systems.

6.3.5. *Simulating Energy Consumption of Wired Networks*

Participant: Anne-Cécile Orgerie.

Predicting the performance of applications, in terms of completion time and resource usage for instance, is critical to appropriately dimension resources that will be allocated to these applications. Current applications, such as web servers and Cloud services, require lots of computing and networking resources. Yet, these resource demands are highly fluctuating over time. Thus, adequately and dynamically dimension these resources is challenging and crucial to guarantee performance and cost-effectiveness. In the same manner, estimating the energy consumption of applications deployed over heterogeneous cloud resources is important in order to provision power resources and make use of renewable energies. Concerning the consumption of entire infrastructures, some studies show that computing resources represent the biggest part in Cloud's consumption, while others show that, depending on the studied scenario, the energy cost of the network infrastructure that links the user to the computing resources can be bigger than the energy cost of the servers. In this work, we aim at simulating the energy consumption of wired networks which receive little attention in the Cloud computing community even though they represent key elements of these distributed architectures. To this end, we are contributing to the well-known open-source simulator ns3 by developing an energy consumption module named ECOFEN.

6.3.6. Simulating the impact of DVFS within SimGrid

Participants: Alexandra Carpen-Amarie, Christine Morin, Anne-Cécile Orgerie.

Simulation is a a popular approach for studying the performance of HPC applications in a variety of scenarios. However, simulators do not typically provide insights on the energy consumption of the simulated platforms. Furthermore, studying the impact of application configuration choices on energy is a difficult task, as not many platforms are equipped with the proper power measurement tools. The goal of this work is to enable energy-aware experimentations within the SimGrid simulation toolkit, by introducing a model of application energy consumption and enabling the use of DVFS techniques for the simulated platforms. We provide the methodology used to obtain accurate energy estimations, highlighting the simulator calibration phase. The proposed energy model is validated by means of a large set of experiments featuring several benchmarks and scientific applications. This work is available in the latest SimGrid release.

6.4. Unconventional Models for Large Computations and Platforms

Participants: Marko Obrovac, Christine Morin, Cédric Tedeschi.

6.4.1. Chemical computing at large scale

Participants: Marko Obrovac, Cédric Tedeschi.

One of the commonly cited problem when dealing with chemistry-inspired computing is its lack of experimental validation. The DHT-based runtime developed recently, in the framework of Marko Obrovac's PhD thesis [13], has been deployed over the Grid'5000 platform with promising results. This runtime is now mature enough for being considered as a viable candidate to underlie a distributed workflow engine [32].

6.4.2. Template workflows

Participants: Christine Morin, Cédric Tedeschi.

In the framework of the DALHIS associate team [5], we plan to combine the high-level template workflow language TIGRES [6], developed by our partner team from Lawrence Berkeley National Lab (LBL) with the workflow management system developed in the team [17]. This work started with the development of a parser of TIGRES.

6.5. Experimental Platforms

Participants: Alexandra Carpen-Amarie, Maxence Dunnewind, Nicolas Lebreton, Julien Lefeuvre, David Margery, Eric Poupart.

[5] http://project.inria.fr/dalhis
[6] http://tigres.lbl.gov/home

6.5.1. Energy measurement

Participants: David Margery, Maxence Dunnewind, Nicolas Lebreton.

In the context of the ECO$_2$Clouds project, the BonFIRE infrastructure was updated. At the hardware level power distribution units that report electricity usage for each outlet were installed. At the software layer, a probe reporting energy sources used was configured. This probe gets its information from RTE, the French Electricity transport network, and allows publication of CO_2 metrics for each machine in the testbed. Moreover, access to these metrics was abstracted through the general API to access BonFIRE.

6.5.2. Deployment of IaaS management system

Participant: Alexandra Carpen-Amarie.

The Grid'5000 platform has become one of the most complete testbeds for designing or evaluating large-scale distributed systems, playing an essential role in enabling experimental research at all levels of the Cloud Computing stack and providing configurable cloud platforms similar to commercially available clouds.

However, the complexity of managing the deployment and tuning of large-scale private clouds emerged as a major drawback. Typically, users study specific cloud components or carry out experiments involving applications running in cloud environments. A key requirement in this context is seamless access to ready-to-use cloud platforms, as well as full control of the deployment settings.

To address these needs, we developed a set of deployment tools for open-source IaaS environments, capable of installing and tuning fully-functional clouds on the Grid'5000 testbed [20]. The deployment tools support four widely-used IaaS clouds, namely OpenNebula, CloudStack, Nimbus and OpenStack.

They rely on the concept of extensible engines for defining experiments. Such engines implement all the stages of an experiment: physical node reservations in Grid'5000, environment deployment, configuration and experiment execution. We designed generic engines for nodes reservation and deployment according to a set of requirements specified in a cloud configuration file. Thus, these engines do not require any prior knowledge of lower-level Grid'5000 tools, allowing the user to easily achieve multi-site Grid'5000 deployments based on multiple environments.

6.5.3. BonFIRE

Participants: Maxence Dunnewind, David Margery, Eric Poupart.

The project was reviewed in December 2013 during CloudCom 2013 in Bristol and rated Excellent. The main achievement this year is the introduction of a reservation system for resources on the BonFIRE platform.

6.5.4. Fed4FIRE

Participants: Julien Lefeuvre, Nicolas Lebreton, David Margery.

In Fed4FIRE, two key technologies have been adopted as common protocols to enable experimenter to interact with testbeds. SFA, to provision resources, and OMF to control them. Here, we contributed to a proposal to secure usage of OMF and to a design to allow using BonFIRE through SFA.

PANAMA Project-Team

6. New Results

6.1. Recent results on sparse representations

Sparse approximation, high dimension, scalable algorithms, dictionary design, sample complexity

The team has had a substantial activity ranging from theoretical results to algorithmic design and software contributions in the field of sparse representations, which is at the core of the ERC project PLEASE (projections, Learning and Sparsity for Efficient Data Processing, see section 8.2.1).

6.1.1. A new framework for sparse representations: analysis sparse models

Participants: Rémi Gribonval, Nancy Bertin, Srdan Kitic, Cagdas Bilen.

Main collaboration: Mike Davies, Mehrdad Yaghoobi (Univ. Edinburgh), Michael Elad (The Technion).

In the past decade there has been a great interest in a synthesis-based model for signals, based on sparse and redundant representations. Such a model assumes that the signal of interest can be composed as a linear combination of *few* columns from a given matrix (the dictionary). An alternative *analysis-based* model can be envisioned, where an analysis operator multiplies the signal, leading to a *cosparse* outcome. Within the SMALL FET-Open project, we initiated a research programme dedicated to this analysis model, in the context of a generic missing data problem (e.g., compressed sensing, inpainting, source separation, etc.). We obtained a uniqueness result for the solution of this problem, based on properties of the analysis operator and the measurement matrix. We also considered a number of pursuit algorithms for solving the missing data problem, including an ℓ^1-based and a new greedy method called GAP (Greedy Analysis Pursuit). Our simulations demonstrated the appeal of the analysis model, and the success of the pursuit techniques presented.

These results have been published in conferences and in a journal paper [19]. Other algorithms based on iterative cosparse projections [83] as well as extensions of GAP to deal with noise and structure in the cosparse representation have been developed, with applications to toy MRI reconstruction problems and acoustic source localization and reconstruction from few measurements [100].

Successful applications of the cosparse approach to sound source localization, audio declipping and brain imaging have been developed this year. In particular, we compared the performance of several cosparse recovery algorithms in the context of sound source localization [39] and showed its efficiency in situations where usual methods fail [60]. It was also shown to be applicable to the hard declipping problem [61]. Application to EEG brain imaging was also investigated and a paper was submitted to ICASSP'14 (see below).

6.1.2. Theoretical results on sparse representations

Participants: Rémi Gribonval, Anthony Bourrier, Pierre Machart.

Main collaboration: Charles Soussen (Centre de recherche en automatique de Nancy (CRAN)), Jérôme Idier (Institut de Recherche en Communications et en Cybernétique de Nantes (IRCCyN)), Cédric Herzet (Equipe-projet FLUMINANCE (Inria - CEMAGREF, Rennes)), Mehrdad Yaghoobi, Mike Davies (University of Edinburgh), Patrick Perez (Technicolor R&I France), Tomer Peleg (The Technion)

Sparse recovery conditions for Orthogonal Least Squares : We pursued our investigation of conditions on an overcomplete dictionary which guarantee that certain ideal sparse decompositions can be recovered by some specific optimization principles / algorithms. We extended Tropp's analysis of Orthogonal Matching Pursuit (OMP) using the Exact Recovery Condition (ERC) to a first exact recovery analysis of Orthogonal Least Squares (OLS). We showed that when ERC is met, OLS is guaranteed to exactly recover the unknown support. Moreover, we provided a closer look at the analysis of both OMP and OLS when ERC is not fulfilled. We showed that there exist dictionaries for which some subsets are never recovered with OMP. This phenomenon, which also appears with ℓ^1 minimization, does not occur for OLS. Finally, numerical experiments based on

our theoretical analysis showed that none of the considered algorithms is uniformly better than the other [21]. More recently, we obtained simpler coherence-based conditions [18] and pursued the analysis of unrecoverable subsets [43].

Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems: The primary challenge in linear inverse problems is to design stable and robust "decoders" to reconstruct high-dimensional vectors from a low-dimensional observation through a linear operator. Sparsity, low-rank, and related assumptions are typically exploited to design decoders which performance is then bounded based on some measure of deviation from the idealized model, typically using a norm. We characterized the fundamental performance limits that can be expected from an ideal decoder given a general model, ie, a general subset of "simple" vectors of interest. First, we extended the so-called notion of instance optimality of a decoder to settings where one only wishes to reconstruct some part of the original high dimensional vector from a low-dimensional observation. This covers practical settings such as medical imaging of a region of interest, or audio source separation when one is only interested in estimating the contribution of a specific instrument to a musical recording. We defined instance optimality relatively to a model much beyond the traditional framework of sparse recovery, and characterized the existence of an instance optimal decoder in terms of joint properties of the model and the considered linear operator [42], [33]. Noiseless and noise-robust settings were both considered [56]. We showed somewhat surprisingly that the existence of noise-aware instance optimal decoders for all noise levels implies the existence of a noise-blind decoder. A consequence of our results is that for models that are rich enough to contain an orthonormal basis, the existence of an L2/L2 instance optimal decoder is only possible when the linear operator is not substantially dimension-reducing. This covers well-known cases (sparse vectors, low-rank matrices) as well as a number of seemingly new situations (structured sparsity and sparse inverse covariance matrices for instance). We exhibit an operator-dependent norm which, under a model-specific generalization of the Restricted Isometry Property (RIP), always yields a feasible instance optimality and implies instance optimality with certain familiar atomic norms such as the ℓ^1 norm.

Connections between sparse approximation and Bayesian estimation: Penalized least squares regression is often used for signal denoising and inverse problems, and is commonly interpreted in a Bayesian framework as a Maximum A Posteriori (MAP) estimator, the penalty function being the negative logarithm of the prior. For example, the widely used quadratic program (with an ℓ^1 penalty) associated to the LASSO / Basis Pursuit Denoising is very often considered as MAP estimation under a Laplacian prior in the context of additive white Gaussian noise (AWGN) reduction.

In 2011 we obtained a result [85] highlighting the fact that, while this is *one* possible Bayesian interpretation, there can be other equally acceptable Bayesian interpretations. Therefore, solving a penalized least squares regression problem with penalty $\phi(x)$ need not be interpreted as assuming a prior $C \cdot \exp(-\phi(x))$ and using the MAP estimator. In particular, we showed that for *any* prior P_X, the minimum mean square error (MMSE) estimator is the solution of a penalized least square problem with some penalty $\phi(x)$, which can be interpreted as the MAP estimator with the prior $C \cdot \exp(-\phi(x))$. Vice-versa, for *certain* penalties $\phi(x)$, the solution of the penalized least squares problem is indeed the MMSE estimator, with a certain prior P_X. In general $dP_X(x) \neq C \cdot \exp(-\phi(x))dx$. This year, we extended this result to general inverse problems [30], [58], [47].

6.1.3. Algorithmic and theoretical results on dictionary learning

Participants: Rémi Gribonval, Nancy Bertin, Cagdas Bilen, Srdan Kitic.

Main collaboration: Rodolphe Jenatton, Francis Bach (Equipe-projet SIERRA (Inria, Paris)), Martin Kleinsteuber, Matthias Seibert (TU-Munich), Mehrdad Yaghoobi, Mike Davies (University of Edinburgh),

Dictionary learning : An important practical problem in sparse modeling is to choose the adequate dictionary to model a class of signals or images of interest. While diverse heuristic techniques have been proposed in the litterature to learn a dictionary from a collection of training samples, there are little existing results which provide an adequate mathematical understanding of the behaviour of these techniques and their ability to recover an ideal dictionary from which the training samples may have been generated.

Beyond our pioneering work [86], [110] [6] on this topic, which concentrated on the noiseless case for non-overcomplete dictionaries, this year we obtained new results showing the relevance of an ℓ^1 penalized cost function for the locally stable identification of overcomplete incoherent dictionaries, in the presence of noise and outliers. Moreover, we established new sample complexity bounds of dictionary learning and other related matrix factorization schemes (including PCA, NMF, structured sparsity ...) [59].

Analysis Operator Learning for Overcomplete Cosparse Representations : Besides standard dictionary learning, we also considered learning in the context of the cosparse model. We consider the problem of learning a low-dimensional signal model from a collection of training samples. The mainstream approach would be to learn an overcomplete dictionary to provide good approximations of the training samples using sparse synthesis coefficients. This famous sparse model has a less well known counterpart, in analysis form, called the cosparse analysis model. In this new model, signals are characterized by their parsimony in a transformed domain using an overcomplete analysis operator.

We considered several approaches to learn an analysis operator from a training corpus [102]. For one of them, which uses a constrained optimization program based on ℓ^1 optimization, we derived a practical learning algorithm, based on projected subgradients, and demonstrated its ability to robustly recover a ground truth analysis operator, provided the training set is of sufficient size. A local optimality condition was derived, providing preliminary theoretical support for the well-posedness of the learning problem under appropriate conditions [24]. Extensions to deal with noisy data have been obtained as well [119].

In more specific situations, when prior information is available on the operator, it is also possible to express the operator on a parametric form, and learn this parameter. For instance, in the sound source localization problem, we showed that unknown speed of sound can be learned jointly in the process of cosparse recovery, under mild conditions. This work was submitted to the iTwist'14 workshop.

6.2. Emerging activities on compressive sensing, learning and inverse problems

Compressive sensing, acoustic wavefields, audio inpainting,

6.2.1. Audio inpainting (SMALL FET-Open project)

Participants: Rémi Gribonval, Nancy Bertin, Corentin Guichaoua, Srdan Kitic.

Inpainting is a particular kind of inverse problems that has been extensively addressed in the recent years in the field of image processing. It consists in reconstructing a set of missing pixels in an image based on the observation of the remaining pixels. Sparse representations have proved to be particularly appropriate to address this problem. However, inpainting audio data has never been defined as such so far.

METISS has initiated a series of works about audio inpainting, from its definition to methods to address it. This research has begun in the framework of the EU Framework 7 FET-Open project FP7-ICT-225913-SMALL (Sparse Models, Algorithms and Learning for Large-Scale data) which began in January 2009. Rémi Gribonval was the coordinator of the project. The research on audio inpainting has been conducted by Valentin Emiya in 2010 and 2011.

The contributions consist of:

- defining audio inpainting as a general scheme where missing audio data must be estimated: it covers a number of existing audio processing tasks that have been addressed separately so far – click removal, declipping, packet loss concealment, unmasking in time-frequency;
- proposing algorithms based on sparse representations for audio inpainting (based on Matching Pursuit and on ℓ^1 minimization);
- addressing the case of audio declipping (*i.e.* desaturation): thanks to the flexibility of our inpainting algorithms, they can be constrained so as to include the structure of signals due to clipping in the objective to optimize. The resulting performance are significantly improved. This work appeared as a journal paper [63].
- addressing the case of audio declipping with the competitive cosparse approach, with promising result especially when the clipping level is low. A contribution was submitted to the iTwist'14 workshop [61].

Current and future works deal with developping advanced sparse decomposition for audio inpainting, including several forms of structured sparsity (*e.g.* temporal and multichannel joint-sparsity), dictionary learning for inpainting, and several applicative scenarios (declipping, time-frequency inpainting).

6.2.2. Blind Calibration of Compressive Sensing systems

Participants: Rémi Gribonval, Cagdas Bilen.

Main collaborations: Gilles Chardon, Laurent Daudet (Institut Langevin), Gilles Puy (EPFL)

We consider the problem of calibrating a compressed sensing measurement system under the assumption that the decalibration consists in unknown gains on each measure. We focus on blind calibration, using measures performed on a few unknown (but sparse) signals. A naive formulation of this blind calibration problem, using ℓ^1 minimization, is reminiscent of blind source separation and dictionary learning, which are known to be highly non-convex and riddled with local minima. In the considered context, when the gains are real valued and non-negative, we showed that in fact this formulation can be exactly expressed as a convex optimization problem, and can be solved using off-the-shelf algorithms. Numerical simulations demonstrated the effectiveness of the approach even for highly uncalibrated measures, when a sufficient number of (unknown, but sparse) calibrating signals is provided. We observed that the success/failure of the approach seems to obey sharp phase transitions [84]. This year, we focused on extending the framework to phase-only decalibration, using techniques revolving around low-rank matrix recovery [27], [26], [34], [52], and to joint phase and gain decalibration [54].

6.2.3. Compressive Gaussian Mixture estimation

Participants: Rémi Gribonval, Anthony Bourrier.

Main collaborations: Patrick Perez (Technicolor R&I France)

When fitting a probability model to voluminous data, memory and computational time can become prohibitive. In this paper, we propose a framework aimed at fitting a mixture of isotropic Gaussians to data vectors by computing a low-dimensional sketch of the data. The sketch represents empirical moments of the underlying probability distribution. Deriving a reconstruction algorithm by analogy with compressive sensing, we experimentally show that it is possible to precisely estimate the mixture parameters provided that the sketch is large enough. Our algorithm provides good reconstruction and scales to higher dimensions than previous probability mixture estimation algorithms, while consuming less memory in the case of numerous data. It also provides a privacy-preserving data analysis tool, since the sketch does not disclose information about individual datum it is based on [38], [40], [29].

6.3. Recent results on tensor decompositions

Multi-linear algebra is defined as the algebra of q-way arrays ($q > 2$), that is, the arrays whose elements are addressed by more than two indices. The first works back as far as Jordan who was interested in simultaneously diagonalizing two matrices at a time [92]. It is noteworthy that such two matrices can be interpreted as both slices of a three-way array and their joint diagonalization can be viewed as Hitchcock's polyadic decomposition [89] of the associated three-way array. Other works followed discussing rank problems related to multi-way structures and properties of multi-way arrays. However, these exercices in multilinear algebra were not linked to real data analysis but stayed within the realm of mathematics. Studying three-way data really started with Tucker's seminal work, which gave birth to the three-mode factor analysis [115]. His model is now often referred to as the Tucker3 model. At the same moment, other authors focused on a particular case of the Tucker3 model, calling it PARAFAC for PARAllel FACtor analysis [88], and on the means to achieve such a decomposition, which will become the famous canonical decomposition [77]. In honor to Hitchcock's pionneer work, we will call it the Canonical Polyadic (CP) decomposition.

Achieving a CP decomposition has been seen first as a mere non-linear least squares problem, with a simple objective criterion. In fact, the objective is a polynomial function of many variables, where some separate. One could think that this kind of objective is easy because smooth, and even infinitely differentiable. But it turns out that things are much more complicated than they may appear to be at first glance. Nevertheless, the Alternating Least Squares (ALS) algorithm has been mostly utilized to address this minimization problem, because of its programming simplicity. This should not hide the inherently complicated theory that lies behind the optimization problem. Moreover, in most of the applications, actual tensors may not exactly satisfy the expected model, so that the problem is eventually an approximation rather than an exact decomposition. This may results in a slow convergence (or lack of convergence) of iterative algorithms such as the ALS one [94]. Consequently, a new class of efficient algorithms able to take into account the properties of tensors to be decomposed is needed.

6.3.1. A novel direct algorithm for CP decompositions

Participant: Laurent Albera.

Main collaborations: Sepideh Hajipour (LTSI & BiSIPL), Isabelle Merlet (LTSI, France), Mohammad Bagher Shamsollahi (BiSIPL, Iran)

Nowadays several techniques are available to solve the CP problem. They can be classified in three main groups [113]: alternating algorithms, which update only a subset of the parameters at each step; derivative-based methods, seeking for an update of all the parameters simultaneously by successive approximations; and direct procedures. The latter algorithms compute the CP decomposition by solving an alternative algebra problem of lower dimensions, but they do not provide a solution in terms of least squares contrarily to the alternating and derivative-based techniques.

We proposed a new direct algorithm to compute the CP decomposition of complex-valued multi-way arrays. The proposed algorithm is based on the Simultaneous Schur Decomposition (SSD) of particular matrices derived from the array to process. We also proposed a new Jacobi-like algorithm to calculate the SSD of several complex-valued matrices. Besides, we analysed our SSD and SSD-based CP techniques in terms of i) identifiability, ii) computational complexity and iii) estimation accuracy through a large number of scenarios including synthetic and real data in the context of CP decomposition. Computer results showed the efficiency of the proposed SSD-based CP method of dealing with some well-known difficult scenarios with swamp-like degeneracies. We also showed that the proposed method outperformed the classical CP algorithms in processing of Paatero multi-way arrays. Finally, the robustness of the proposed algorithm with respect to overfactoring was highlighted. This work was briefly presented at ICASSP'13 [31] while a journal paper for submission to IEEE Transactions on Signal Processing is in preparation.

6.3.2. CP decomposition of semi-symmetric semi-nonnegative three-way arrays

Participant: Laurent Albera.

Main collaboration (line search methods): Julie Coloigner (LTSI, France), Amar Kachenoura (LTSI, France), Lotfi Senhadji (LTSI, France)

Main collaborations (Jacobi-like approaches): Lu Wang (LTSI, France), Amar Kachenoura (LTSI, France), Lotfi Senhadji (LTSI, France), Huazhong Shu (LIST, China)

We proposed new algorithms for the CP decomposition of semi-nonnegative semi-symmetric three-way tensors. In fact, it consists in fitting the CP model for which two of the three loading matrices are nonnegative and equal. Note that such a problem can also be interpreted as a nonnegative Joint Diagonalization by Congruence (JDC) problem.

Line search and trust region strategies

We first circumvented the nonnegativity constraint by means of changes of variable into squares, leading to a (polynomial) unconstrained optimization problem. Two optimization strategies, namely line search and trust region, were then studied. Regarding the former, a global plane search scheme was considered. It consists in computing, for a given direction, one or two optimal stepsizes, depending on whether the same stepsize is used in various updating rules. Moreover, we provided a compact matrix form for the derivatives of the objective function. This allows for a direct implementation of several iterative algorithms such as Conjugate Gradient (CG), Levenberg-Marquardt (LM) and Newton-like methods, in matrix programming environments like MATLAB. Note that the computational complexity issue was taken into account in the design phase of the algorithms, and was evaluated for each algorithm, allowing to fairly compare their performance.

Thus, various scenarios have been considered, aiming at testing the influence of i) an additive noise, which can stand for modeling errors, ii) the collinearity between factors, iii) the array rank and iv) the data size. The comparisons between our CG-like, Newton-like and LM-like methods (where semi-nonnegativity and semi-symmetry constraints are exploited), and classical CP algorithms (where no constraints are considered), showed that a better CP decomposition is obtained when these a priori are exploited, especially in the context of high dimensions and high collinearity. Finally, based on our numerical analysis, the algorithms that seem to yield the best tradeoff between accuracy and complexity are our $\mathrm{CG}_{2\mathrm{steps}}$-like and LM-like algorithms.

This work was accepted for publication with minor revisions to the Elsevier Linear Algebra and Applications journal.

Next, we considered an exponential change of variable leading to a different (non-polynomial) unconstrained optimization problem. Then we proposed novel algorithms based on line search strategy with an analytic global plane search procedure requiring new matrix derivations. Their performance was evaluated in terms of estimation accuracy and computational complexity. The classical ELS-ALS [109] and LM [113] algorithms without symmetry and nonnegativity constraints, and the ACDC algorithm [120] where only the semi-symmetry constraint is imposed, were tested as reference methods. Furthermore, the performance was also compared with our algorithms based on a square change of variable. The comparison studies showed that, among these approaches, the best accuracy/complexity trade off was achieved when an exponential change of variable was used through our ELS-ALS-like algorithm.

This work was submitted to the Elsevier Signal Processing journal.

Jacobi-like approaches

The line search (despite the use of global plane search procedures) and trust region strategies may be sensitive to initialization, and generally require a multi-initialization procedure. In order to circumvent this drawback, we considered in this work Jacobi-like approaches, which are known to be less sensitive to initialization. Note that our line search and trust region approaches can then be used to refine the solution obtained by the latter.

More particularly, we formulated the high-dimensional optimization problem into several sequential polynomial subproblems using i) a square change of variables to impose nonnegativity and ii) LU matrix factorization for parameterization. The two equal nonnegative loading matrices are actually written as the Hadamard product of two equal matrices which can be factorized as the product of elementary lower and upper triangular matrices, each one depending on only one parameter.

The first approach minimizes alternatively the classical least squares objective criterion with respect to each parameter of the two equal nonnegative loading matrices and each column of the third loading matrix. This work was published in the IEEE Signal Processing Letters journal [23]. The second technique reduces the previous optimization problem to the computation of the two equal nonnegative loading matrices. The third loading matrix is algebraically derived from the latter. This requires an appropriate parameterization of the set of matrices whose inverse is nonnegative. This work was briefly presented at EUSIPCO'13 [37] while a journal paper for submission to IEEE Transactions on Signal Processing is in preparation. Numerical experiments on simulated matrices emphasize the advantages of the proposed algorithms over classical CP and JDC techniques, especially in the case of degeneracies.

6.4. Source separation and localization

Source separation, sparse representations, tensor decompositions, semi-nonnegative independent component analysis, probabilistic model, source localization

6.4.1. A general framework for audio source separation

Participants: Frédéric Bimbot, Rémi Gribonval, Nancy Bertin.

Main collaboration: E. Vincent (EPI PAROLE, Inria Nancy); N.Q.K. Duong (Technicolor R&I France)

Source separation is the task of retrieving the source signals underlying a multichannel mixture signal. The state-of-the-art approach consists of representing the signals in the time-frequency domain and estimating the source coefficients by sparse decomposition in that basis. This approach relies on spatial cues, which are often not sufficient to discriminate the sources unambiguously. Recently, we proposed a general probabilistic framework for the joint exploitation of spatial and spectral cues [103], which generalizes a number of existing techniques including our former study on spectral GMMs [66]. This framework makes it possible to quickly design a new model adapted to the data at hand and estimate its parameters via the EM algorithm. As such, it is expected to become the basis for a number of works in the field, including our own.

Since the EM algorithm is sensitive to initialization, we devoted a major part of our work to reducing this sensitivity. One approach is to use some prior knowledge about the source spatial covariance matrices, either via probabilistic priors [82] or via deterministic subspace constraints [91]. The latter approach was the topic of the PhD thesis of Nobutaka Ito [90]. A complementary approach is to initialize the parameters in a suitable way using source localization techniques specifically designed for environments involving multiple sources and possibly background noise [74]. This year, we showed that the approach provides a statistically principled solution to the permutation problem in a semi-infomed scenario where the source positions and certain room characteristics are known [15].

6.4.2. Towards real-world separation and remixing applications

Participants: Nancy Bertin, Frédéric Bimbot, Jules Espiau de Lamaestre, Jérémy Paret, Laurent Simon, Nathan Souviraà-Labastie, Joachim Thiemann.

Shoko Araki, Jonathan Le Roux (NTT Communication Science Laboratories, JP), E. Vincent (EPI PAROLE, Inria Nancy)

Following our founding role in the organization of the Signal Separation Evaluation Campaigns (SiSEC) [65], [101], our invited paper summarized the outcomes of the three first editions of this campaign from 2007 to 2010 [116]. While some challenges remain, this paper highlighted that progress has been made and that audio source separation is closer than ever to successful industrial applications. This is also exemplified by the ongoing i3DMusic project and the contracts with Canon Research Centre France and MAIA Studio.

Our involvement in evaluation campaigns and source separation community was reinforced by the recording and the public release of the DEMAND (Diverse Environments Multi-channel Acoustic Noise Database) database, which provides multichannel real-world indoor and outdoor environment noise [44] under Creative Commons licence.

In order to exploit our know-how for these real-world applications, we investigated issues such as how to implement our algorithms in real time [111], how to adapt EM rules for faster computation in multichannel setting [35], how to reduce artifacts [96], how our techniques compare to beamforming in realistic conditions [36], and (in the context of our collaboration with MAIA studios) how best to exploit extra information or human input. In addition, while the state-of-the-art quality metrics previously developed by METISS remain widely used in the community, we proposed some improvements to the perceptually motivated metrics introduced last year [117].

6.4.3. Exploiting filter sparsity for source localization and/or separation

Participants: Alexis Benichoux, Rémi Gribonval, Frédéric Bimbot.

E. Vincent (EPI PAROLE, Inria Nancy)

Estimating the filters associated to room impulse responses between a source and a microphone is a recurrent problem with applications such as source separation, localization and remixing.

We considered the estimation of multiple room impulse responses from the simultaneous recording of several known sources. Existing techniques were restricted to the case where the number of sources is at most equal to the number of sensors. We relaxed this assumption in the case where the sources are known. To this aim, we proposed statistical models of the filters associated with convex log-likelihoods, and we proposed a convex optimization algorithm to solve the inverse problem with the resulting penalties. We provided a comparison between penalties via a set of experiments which shows that our method allows to speed up the recording process with a controlled quality tradeoff [72], [71]. This was a central part of the Ph.D. thesis of Alexis Benichoux [12] defended this year. A journal paper including extensive experiments with real data has been submitted [69].

We also investigated the filter estimation problem in a blind setting, where the source signals are unknown. On a more theoretical side, we studied the frequency permutation ambiguity traditionnally incurred by blind convolutive source separation methods. We focussed on the filter permutation problem in the absence of scaling, investigating the possible use of the temporal sparsity of the filters as a property enabling permutation correction. The obtained theoretical and experimental results highlight the potential as well as the limits of sparsity as an hypothesis to obtain a well-posed permutation problem. This work has been published in a conference [70] and as a journal paper [14].

Finally, we considered the problem of blind sparse deconvolution, which is common in both image and signal processing. To counter-balance the ill-posedness of the problem, many approaches are based on the minimization of a cost function. A well-known issue is a tendency to converge to an undesirable trivial solution. Besides domain specific explanations (such as the nature of the spectrum of the blurring filter in image processing) a widespread intuition behind this phenomenon is related to scaling issues and the nonconvexity of the optimized cost function. We proved that a fundamental issue lies in fact in the intrinsic properties of the cost function itself: for a large family of shift-invariant cost functions promoting the sparsity of either the filter or the source, the only global minima are trivial. We completed the analysis with an empirical method to verify the existence of more useful local minima [25].

6.4.4. Semi-nonnegative independent component analysis

Participant: Laurent Albera.

Main collaborations: Lu Wang (LTSI, France), Amar Kachenoura (LTSI, France), Lotfi Senhadji (LTSI, France), Huazhong Shu (LIST, China)

Independent Component Analysis (ICA) plays an important role in many areas including biomedical engineering [93], [64], [95], [118], [106], [81], speech and audio [67], [68], [78], [75], radiocommunications [80] and document restoration [114] to cite a few.

For instance in [114], the authors use ICA to restore digital document images in order to improve the text legibility. Indeed, under the statistical independence assumption, authors succeed in separating foreground text and bleed-through/show-through in palimpsest images. Furthermore, authors in [81] use ICA to solve the ambiguity in X-ray images due to multi-object overlappings. They presented a novel object decomposition technique based on multi-energy plane radiographs. This technique selectively enhances an object that is characterized by a specific chemical composition ratio of basis materials while suppressing the other overlapping objects. Besides, in the context of classification of tissues and more particularly of brain tumors [106], ICA is very effective. In fact, it allows for feature extraction from Magnetic Resonance Spectroscopy (MRS) signals, representing them as a linear combination of tissue spectra, which are as independent as possible [112]. Moreover, using the JADE algorithm [76] applied to a mixture of sound waves computed by means of the constant-Q transform (Fourier transform with log-frequency) of a temporal waveform broken up into a set of time segments, the authors of [75] describe trills as a set of note pairs described by their spectra and corresponding time envelopes. In this case, pitch and timing of each note present in the trill can be easily deduced.

All the aforementioned applications show the high efficiency of the ICA and its robustness to the presence of noise. Despite this high efficiency in resolving the proposed applicative problems, authors did not fully exploit properties enjoyed by the mixing matrix such as its nonnegativity. For instance in [81], the thickness of each organ, which stands for the mixing coefficient, is real positive. Furthermore, reflectance indices in [114] for the background, the overwriting and the underwriting, which correspond to the mixing coefficients, are also nonnegative. Regarding tissue classification from MRS data, each observation is a linear combination of independent spectra with positive weights representing concentrations [87]; the mixing matrix is again nonnegative.

By imposing the nonnegativity of the mixing matrix within the ICA process, we shown through computer results that the extraction quality can be improved. Exploiting the nonnegativity property of the mixing matrix during the ICA process gives rise to what we call semi-nonnegative ICA. More particularly, we performed the latter by computing a constrained joint CP decomposition of cumulant arrays of different orders [98] having the nonnegative mixing matrix as loading matrices. After merging the entries of the cumulant arrays in the same third order array, the reformulated problem follows the semi-symmetric semi-nonnegative CP model defined in section 6.3.2 . Hence we use the new methods described in section 6.3.2 to perform semi-nonnegative ICA. Performance results in audio and biomedical engineering were given in the different papers cited in section 6.3.2 .

6.4.5. Brain source localization

Participants: Laurent Albera, Srdan Kitic, Nancy Bertin, Rémi Gribonval.

Main collaborations: Hanna Becker (GIPSA & LTSI, France), Isabelle Merlet (LTSI, France), Fabrice Wendling (LTSI, France), Pierre Comon (GIPSA, France), Christian Benar (La Timone, Marseille), Martine Gavaret (La Timone, Marseille), Gwenaël Birot (FBML, Genève), Martin Haardt (TUI, Germany)

Main collaborations: Hanna Becker (GIPSA & LTSI, France), Pierre Comon (GIPSA, France), Isabelle Merlet (LTSI, France), Fabrice Wendling (LTSI, France)

Tensor-based approaches

The localization of several simultaneously active brain regions having low signal-to-noise ratios is a difficult task. To do this, tensor-based preprocessing can be applied, which consists in constructing a Space-Time-Frequency (STF) or Space-Time-Wave-Vector (STWV) tensor and decomposing it using the CP decomposition. We proposed a new algorithm for the accurate localization of extended sources based on the results of the tensor decomposition. Furthermore, we conducted a detailed study of the tensor-based preprocessing methods, including an analysis of their theoretical foundation, their computational complexity, and their performance for realistic simulated data in comparison to three conventional source localization algorithms, namely sLORETA [105], cortical LORETA (cLORETA) [104], and 4-ExSo-MUSIC [73]. Our objective consisted, on the one hand, in demonstrating the gain in performance that can be achieved by tensor-based preprocessing, and, on the other hand, in pointing out the limits and drawbacks of this method. Finally, we validated the STF and STWV techniques on real epileptic measurements to demonstrate their usefulness for practical applications. This work was recently submitted to the Elesevier NeuroImage journal.

From tensor to sparse models

The brain source imaging problem has been widely studied during the last decades, giving rise to an impressive number of methods using different priors. Nevertheless, a thorough study of the latter, including especially sparse and tensor-based approaches, is still missing. Consequently, we proposed i) a taxonomy of the methods based on a priori assumptions, ii) a detailed description of representative algorithms, iii) a review of identifiability results and convergence properties of different techniques, and iv) a performance comparison of the selected methods on identical data sets. Our aim was to provide a reference study in the biomedical engineering domain which may also be of interest for other areas such as wireless communications, audio source localization, and image processing where ill-posed linear inverse problems are encountered and to identify promising directions for future research in this area. A part of this work was submitted to ICASSP'14 while the whole part was submitted to IEEE Signal Processing Magazine.

A cosparsity-based approach

Cosparse modeling is particularly attractive when the signals of interest satisfy certain physical laws that naturally drive the choice of an analysis operator. We showed how to derive a reduced non-singular analysis operator describing EEG signals from Poisson's equation, Kirchhoff's law and some other physical constraints. As a result, we proposed the CoRE (Cosparse Representation of EEG signals) method to solve the classical brain source imaging problem. Computer simulations demonstrated the numerical performance of the CoRE method in comparison to a dictionary-based sparse approach. This work was submitted to ICASSP'14.

6.5. Audio and speech content processing

Audio segmentation, speech recognition, motif discovery, audio mining

6.5.1. Audio motif discovery

Participants: Frédéric Bimbot, Laurence Catanese.

This work was performed in close collaboration with Guillaume Gravier from the Texmex project-team.

As an alternative to supervised approaches for multimedia content analysis, where predefined concepts are searched for in the data, we investigate content discovery approaches where knowledge emerge from the data. Following this general philosophy, we pursued work on motif discovery in audio contents.

Audio motif discovery is the task of finding out, without any prior knowledge, all pieces of signals that repeat, eventually allowing variability. The developed algorithms allows discovering and collecting occurrences of repeating patterns in the absence of prior acoustic and linguistic knowledge, or training material.

Former work extended the principles of seeded discovery to near duplicate detection and spoken document retrieval from examples [99].

In 2012, the work achieved consisted in consolidating previously obtained results with the motif discovery algorithm and making implementation choices regardless of the structure and the code, in order to minimize the computation time. This has lead to the creation of a software prototype called MODIS.

After the code has been thoroughly optimised, further optimizations to improve the system performances was to change the method used for the search of similarities between patterns. A new functionality has been added to get rid of unrelevant patterns like silence in speech. New versions of dynamic time warping have been implemented, as well as the possibility to downsample the input sequence during the process, which allows a huge gain of computation time.

The principles of the MODIS software has been documented in details [48] and demonstrated during a Show & Tell session at the Interspeech 2013 conference [41].

This work has been carried out in the context of the Quaero Project.

6.5.2. Landmark-driven speech recognition

Participant: Stefan Ziegler.

This work is supervised by Guillaume Gravier and Bogdan Ludusan from the Texmex project-team.

Our previous studies indicate that acoustic-phonetic approaches to ASR, while they cannot achieve state-of-the-art ASR performance by themselves, can prevent HMM-based ASR from degrading, by integrating additional knowledge into the decoding.

In our previous framework we inserted knowledge into the decoding by detecting time frames (referred to as landmarks) which estimate the presence of the active broad phonetic class. This enables the use of a modified version of the viterbi decoding that favours states that are coherent with the detected phonetic knowledge [122].

In 2012 we focused on two major issues. First, we aimed at finding new ways to model and detect phonetic landmarks. Our second focus was on the extension of our landmark detector towards a full acoustic-phonetic framework, to model speech by a variety of articulatory features.

Our new approach for the classification and detection of speech units focuses on developping landmark-models that are different from existing frame-based approaches to landmark detection [121]. In our approach, we use segmentation to model any time-variable speech unit by a fixed-dimensional observation vector. After training any desired classifier, we can estimate the presence of a desired speech unit by searching for each time frame the corresponding segment, that provides the maximum classification score.

We used this segment-based landmark-detection inside a standalone acoustic-phonetic framework that models speech as a stream of articulatory features. In this framework we first search for relevant broad phonetic landmarks, before attaching each landmark with the full set of articulatory features.

Integrating these articulatory feature streams into a standard HMM-based speech recognizer by weighted linear combination improves speech recognition up to 1.5

Additionally, we explored the possibilities of using stressed syllables as an information to guide the viterbi decoding. This work was carried under the leaderhip of Bogdan Ludusan from the team TEXMEX at IRISA [97].

6.5.3. Mobile device for the assistance of users in potentially dangerous situations

Participants: Romain Lebarbenchon, Frédéric Bimbot.

The S-Pod project is a cooperative project between industry and academia aiming at the development of mobile systems for the detection of potentially dangerous situations in the immediate environment of a user, without requiring his/her active intervention.

In this context, the PANAMA research group is involved in the design of algorithms for the analysis and monitoring of the acoustic scene around the user, yielding information which can be fused with other sources of information (physiological, contextual, etc...) in order to trigger an alarm when needed and subsequent appropriate measures.

Currently in its initial phase, work has mainly focused on functional specifications and performance requirements.

6.6. Music Content Processing and Music Information Retrieval

Acoustic modeling, non-negative matrix factorisation, music language modeling, music structure

6.6.1. Music language modeling

Participants: Frédéric Bimbot, Dimitri Moreau, Stanislaw Raczynski.

Main collaboration: S. Fukayama (University of Tokyo, JP), E. Vincent (EPI PAROLE, Inria Nancy), Intern: A. Aras

Music involves several levels of information, from the acoustic signal up to cognitive quantities such as composer style or key, through mid-level quantities such as a musical score or a sequence of chords. The dependencies between mid-level and lower- or higher-level information can be represented through acoustic models and language models, respectively.

We pursued our pioneering work on music language modeling, with a particular focus on the joint modeling of "horizontal" (sequential) and "vertical" (simultaneous) dependencies between notes by log-linear interpolation of the corresponding conditional distributions. We identified the normalization of the resulting distribution as a crucial problem for the performance of the model and proposed an exact solution to this problem [108]. We also applied the log-linear interpolation paradigm to the joint modeling of melody, key and chords, which evolve according to different timelines [107]. In order to synchronize these feature sequences, we explored the use of beat-long templates consisting of several notes as opposed to short time frames containing a fragment of a single note.

The limited availability of multi-feature symbolic music data is currently an issue which prevents the training of the developed models on sufficient amounts of data for the unsupervised probabilistic approach to significantly outperform more conventional approaches based on musicological expertise. We outlined a procedure for the semi-automated collection of large-scale multifeature music corpora by exploiting the wealth of music data available on the web (audio, MIDI, leadsheets, lyrics, etc) together with algorithms for the automatic detection and alignment of matching data. Following this work, we started collecting pointers to data and developing such algorithms.

Effort was dedicated to the investigation of structural models for improving the modeling of chord sequence. Preliminary results obtained during Anwaya Aras' intersnship show that using a matricial structure of time dependencies between successive chords improves the predictability of chord sequences as compared to a purely sequential model.

6.6.2. Music structuring

Participants: Frédéric Bimbot, Anaik Olivero, Gabriel Sargent.

Main collaboration: E. Vincent (EPI PAROLE, Inria Nancy), Intern: E. Deruty

The structure of a music piece is a concept which is often referred to in various areas of music sciences and technologies, but for which there is no commonly agreed definition. This raises a methodological issue in MIR, when designing and evaluating automatic structure inference algorithms. It also strongly limits the possibility to produce consistent large-scale annotation datasets in a cooperative manner.

Last year, our methodology for the *semiotic* annotation of music pieces has developed and concretized into a set of principles, concepts and conventions for locating the boundaries and determining metaphoric labels of music segments. The method relies on a new concept for characterizing the inner organization of music segments called the System & Contrast (S&C) model [2]. The annotation of 383 music pieces has been finalized, documented [28] and released to the MIR scientific community: http://musicdata.gforge.inria.fr/structureAnnotation.html.

For what concerns algorithmic approaches to music structure description [13], we have formulated the segmentation process as the optimization of a cost function which is composed of two terms: the first one corresponds to the characterization of structural segments by means of audio criteria; the second one relies on the regularity of the target structure with respect to a "structural pulsation period". In this context, we have compared several regularity constraints and studied the combination of audio criteria through fusion. We also considered the estimation of structural labels as a probabilistic finite-state automaton selection process : in this scope, we have proposed an auto-adaptive criterion for model selection, applied to a description of the tonal content. We also proposed a labeling method derived from the system-contrast model. We have evaluated and compared several systems for structural segmentation of music based on these approaches in the context of national and international evaluation campaigns (Quaero, MIREX).

As a follow-up to this work on music structure description, we are currently designing new models and algorithms for segmenting and labeling music into structural units. In one approach (Corentin Guichaoua's PhD), music structure is described as a hierarchical tree estimated by a grammar inference process whereas a second approach (Anaik Olivero's Post-doc) addresses music structure description as the estimation of a graph of similarity relationships.

S4 Project-Team

6. New Results

6.1. New result 1

SAGE Project-Team

6. New Results

6.1. Parallel numerical algorithms

6.1.1. Parallel Adaptive GMRES with deflated restarting

Participant: Jocelyne Erhel.

Grants and projects: C2S@EXA 8.2.3 , JLPC 8.4.4

Software: DGMRES, AGMRES, GPREMS.

Publications: [17], [26].

Abstract: The GMRES iterative method is widely used as a Krylov subspace technique for solving sparse linear systems when the coefficient matrix is nonsymmetric and indefinite. The Newton basis implementation has been proposed on distributed memory computers as an alternative to the classical approach with the Arnoldi process. The aim of our work here is to introduce a modification based on deflation techniques. This approach builds an augmented subspace in an adaptive way to accelerate the convergence of the restarted formulation. In our numerical experiments, we show the benefits of using this implementation with hybrid direct/iterative methods to solve large linear systems.

6.1.2. Hybrid algebraic solvers for CFD problems

Participant: Jocelyne Erhel.

Grants and projects: C2S@EXA 8.2.3 , JLPC 8.4.4

Software: DGMRES, AGMRES, GPREMS.

Publications: [18].

Abstract: Sparse linear systems arise from design optimization in computational fluid dynamics. In this approach, a linearization of the discretized compressible Navier-Stokes equations is built, in order to evaluate the sensitivity of the entire flow with respect to each design parameter. The goal is to reduce the memory requirements and indirectly, the computational cost at different steps of this scheme. Numerical results are presented with industrial test cases to show the benefits of our methodology.

6.1.3. Algebraic multilevel preconditioning

Participant: Thomas Dufaud.

Grants: C2S@EXA 8.2.3

Publications: [51], [23], [24].

Conferences: [37], [24].

Abstract: The Schwarz domain decomposition method is a very attractive numerical method for parallel computing as it needs only to update the boundary conditions on the artificial interfaces generated by domain decomposition. Thus only local communications between the neighbouring sub-domains are required. We review the use of Aitken's acceleration applied to the Schwarz domain decomposition method.

6.1.4. Counting eigenvalues in domains of the complex field

Participant: Bernard Philippe.

Grants: momappli 8.4.2

Publications: [15], [28].

Conferences: [47], [48], [22].

Abstract: A procedure for counting the number of eigenvalues of a matrix in a region surrounded by a closed curve is presented. It is based on the application of the residual theorem. The quadrature is performed by evaluating the principal argument of the logarithm of a function. A strategy is proposed for selecting a path length that insures that the same branch of the logarithm is followed during the integration. Numerical tests are reported for matrices obtained from conventional matrix test sets.

6.1.5. Sliced-time computation method

Participant: Jocelyne Erhel.

Grants: MODNUM 8.4.5

Publications: [16], [25].

Abstract: We consider the mathematical framework of a sliced-time computation method for explosive solutions to systems of ordinary differential equations. We also derive an Adaptive Parallel-in-Time Method with application to a membrane problem.

6.1.6. Interacting particles systems

Participant: Lionel Lenôtre.

Grants: H2MNO4 8.2.1

Conferences: [31]

Abstract: We consider a variance reduction method for simulations with particles.

6.2. Numerical models and simulations applied to physics

6.2.1. Heat and mass transfer modeling in porous media

Participants: Édouard Canot, Salwa Mansour.

Grants: MODNUM 8.4.5 , HYDRINV 8.4.7

Conferences: [33], [35],[44]

Abstract: The effective thermal conductivity is a key parameter for obtaining good simulations of heat transfer in wet porous media. It is very sensitive to the presence of liquid water, even in very small quantity. Moreover, during the evaporation of water, some changes of geometric configuration of the liquid meniscus lead to hysteresis behaviors. Micro-scale studies help us in understanding the global properties, via numerical simulations.

6.2.2. Heat transfer in soils applied to archaeological fires

Participants: Édouard Canot, Salwa Mansour.

Grants: MODNUM 8.4.5 , ARPHYMAT 8.4.6

Conferences: [34], [36]

Abstract: In order to be validated, the numerical simulations of heat transfer at the surface of the soil are compared to experimental results, because of the complexity of the phenomenon and the great number of physical mechanisms involved. It appears that making good experiments is hard, not to mention the limitations and lacks of the Laloy and Massard method used to obtained the effective thermal conductivity of the granular material. The Laloy and Massard method have been slightly improved; besides a different, new experimental method, based on the mathematical properties of heat transfer, has been proposed.

6.2.3. Granular materials

Participant: Édouard Canot.

Publications: [19].

Abstract: Using the $\mu(I)$ continuum model recently proposed for dense granular flows, we study theoretically steady and fully developed granular flows in two configurations: a plane shear cell and a channel made of two parallel plates (Poiseuille configuration).

6.2.4. Geodesy

Participants: Amine Abdelmoula, Bernard Philippe.

Grants: LIRIMA-EPIC 8.4.3 , joint Ph-D 8.4.9 .

Publications: [12].

Thesis: Ph-D of Amine Abdelmoula, University of Rennes 1 and Tunis, defended in December 2013.

Abstract: We solve a geodetic inverse problem for the determination of a distribution of point masses (characterized by their intensities and positions), such that the potential generated by them best approximates a given potential field.

6.3. Models and simulations for flow and transport in porous media

6.3.1. Flow and transport in highly heterogeneous porous medium

Participants: Jean-Raynald de Dreuzy, Jocelyne Erhel, Géraldine Pichot.

Grants: H2MN04 8.2.1 , H2OGuilde 8.2.4 , HEMERA 8.2.2

Software: PARADIS, H2OLab

Publications: [13]

Abstract: Models of hydrogeology must deal with both heterogeneity and lack of data. We consider a flow and transport model for an inert solute. The conductivity is a random field following a stationary log normal distribution with an exponential or Gaussian covariance function, with a very small correlation length. The quantities of interest studied here are the expectation of the spatial mean velocity, the equivalent permeability and the macro spreading. In particular, the asymptotic behavior of the plume is characterized, leading to large simulation times, consequently to large physical domains. Uncertainty is dealt with a classical Monte Carlo method, which turns out to be very efficient, thanks to the ergodicity of the conductivity field and to the very large domain. These large scale simulations are achieved by means of high performance computing algorithms and tools.

6.3.2. Diffusion processes in porous media

Participants: Lionel Lenôtre, Géraldine Pichot.

Grants: H2MN04 8.2.1

Software: SBM 5.1.7 , PALMTREE 5.2.1

Publications: [21]

Conferences: [41], [43], [42]

Abstract: We present some recent results about Monte Carlo simulations in media with interfaces. By nature, porous media are extremely heterogeneous. We consider a one-dimensional advection-diffusion equation with piecewise constant coefficients. Without drift term, the Skew Brownian Motion permits to develop several exact algorithms with constant time step. We aim at adding the drift term and dealing with higher dimensional problems.

6.3.3. Adaptive stochastic model for flow and transport with random data

Participants: Jocelyne Erhel, Mestapha Oumouni.

Grants: HYDRINV 8.4.7 , joint Ph-D 8.4.8

Publications: [27].

Conferences: [46].

Thesis:[11].

Abstract: This work presents a development and an analysis of an effective approach for partial differential equation with random coefficients and data. We are interesting in the steady flow equation with stochastic input data.

A projection method in the one-dimensional case is presented to compute efficiently the average of the solution.

An anisotropic sparse grid collocation method is also used to solve the flow problem. First, we introduce an indicator of the error satisfying an upper bound of the error, it allows us to compute the anisotropy weights of the method. We demonstrate an improvement of the error estimation of the method which confirms the efficiency of the method compared with Monte Carlo and will be used to accelerate this method by the Richardson extrapolation technique.

We also present a numerical analysis of a probabilistic method to quantify the migration of a contaminant in random media. We consider the previous flow problem coupled with the advection-diffusion equation, where we are interested in the computation of the mean extension and the mean dispersion of the solute. The flow model is discretized by a mixed finite elements method and the concentration of the solute is the density of the solution of a stochastic differential equation, which is discretized by an Euler scheme. We present an explicit formula of the dispersion and optimal a priori error estimates.

6.3.4. Reactive transport

Participants: Édouard Canot, Jocelyne Erhel, Souhila Sabit.

Grants: H2MN04 8.2.1 , ANDRA 7.1 , MOMAS 8.2.7 , C2SEXA 8.2.3

Software: GRT3D.

Publications: [52],[30].

Conferences: [20],[40],[50], [32].

Abstract: Numerical simulations are essential for studying the fate of contaminants in aquifers, for risk assessment and resources management. In this study, we deal with reactive transport models and show how a Newton method can be used efficiently. Numerical experiments illustrate the efficiency of a substitution technique. Moreover, it appears that using logarithms in the chemistry equations lead to ill conditioned matrices and increase the computational cost.

6.4. Models and simulations for flow in porous fractured media

6.4.1. Synthetic benchmark for modeling flow in 3D fractured media

Participants: Jean-Raynald de Dreuzy, Jocelyne Erhel, Géraldine Pichot.

Grants: GEOFRAC 8.2.5 , FRACINI 8.1.1

Software: MPFRAC

Publications: [14]

Abstract: Intensity and localization of flows in fractured media have promoted the development of a large range of different modeling approaches including Discrete Fracture Networks, pipe networks and equivalent continuous media. While benchmarked usually within site studies, we propose an alternative numerical benchmark based on highly-resolved Discrete Fracture Networks (DFNs) and on a stochastic approach. Test cases are built on fractures of different lengths, orientations, aspect ratios and hydraulic apertures, issuing the broad ranges of topological structures and hydraulic properties classically observed. We present 18 DFN cases, with 10 random simulations by case.

6.4.2. Robust numerical methods for solving flow in stochastic fracture networks

Participants: Jean-Raynald de Dreuzy, Jocelyne Erhel, Géraldine Pichot.

Grants: GEOFRAC 8.2.5 , FRACINI 8.1.1

Software: MPFRAC, H2OLab.

Publications: [29].

Conferences: [49].

Abstract: In this work, flow in Discrete Fracture Networks (DFN) is solved using a Mortar Mixed Hybrid Finite Element Method. To solve large linear systems derived from a nonconforming discretization of stochastic fractured networks, a Balancing Domain Decomposition is used. Tests on three stochastically generated DFN are proposed to show the ability of the iterative solver SIDNUR to solve the flow problem.

6.4.3. Flow in complex 3D geological fractured porous media

Participants: Jean-Raynald de Dreuzy, Thomas Dufaud, Jocelyne Erhel, Géraldine Pichot.

Grants: GEOFRAC 8.2.5 , FRACINI 8.1.1

Software: MPFRAC, H2OLab

Conferences: [38], [39]

Abstract: Taking into account water and solute exchanges between porous and fractured media is of great interest in geological applications. The coupled porous-fractured flow equations and their discretization by a Mixed Hybrid Finite Element Method are presented as well as the derived linear system. An appropriate mesh generation is proposed to deal with the complexity involved by randomly generated fracture networks. Numerical experiments are shown, that provide flow fields for forthcoming transport simulations.

SERPICO Project-Team

6. New Results

6.1. Lifetime estimation in photon counting-based fluorescence lifetime imaging microscopy

Participants: Philippe Roudot, Charles Kervrann.

In this study, we investigated a Maximum Likelihood (ML) framework for photon counting-based fluorescence lifetime estimation in Fluorescence Lifetime Imaging Microscopy (FLIM). Data collected at a given pixel consist of photon counts exponentially decreasing along the time and are assumed to follow Poisson statistics (see Fig. 6). A careful analysis of the biophysical phenomenon and instrument models are used to derive a proper ML framework for lifetime estimation. Unlike usual pointwise approaches, a neighborhood-wise approach is proposed to take explicitly into account the spatial correlation of data [15]. The application to real biological data allowed us to prove the spatial localisation of interactions, a new result which was not achievable with conventional methods. For future work, the main challenge is to extend the framework to deal with multi-exponential decay estimate and adaptive neighborhoods, a challenge we need to address for a large class of biological studies.
Reference: [15]
Partners: A. Chessel (University of Cambridge, UK), F. Waharte and J. Boulanger (UMR 144, PICT IBiSA, CNRS-Institut Curie)

6.2. Vesicle segmentation method with automatic scale selection in TIRF microscopy

Participants: Antoine Basset, Charles Kervrann, Patrick Bouthemy.

Accurately detecting cellular structures in fluorescence microscopy is of primary interest for further quantitative analysis such as counting, tracking or classification. We aimed at segmenting vesicles in Total Internal Reflection Fluorescence (TIRF) microscopy images.

In this study, we have proposed an original and efficient method – called SLT-LoG – for vesicle segmentation with fewer parameters than the state-of-the-art methods. It exploits the Laplacian of Gaussian (LoG) of the images at several scales. Since the vesicles size is almost constant in space and time, a prominent mode is expected in the empirical distribution of the scales at which the minima of LoG values are detected. It precisely corresponds to the optimal sought scale. The vesicle segmentation map is then derived by thresholding the LoG values obtained at this optimal scale. To set the threshold, we assume that the values of the LoG locally follow a normal distribution (see Fig. 7). For each point, we estimate the local mean and variance, and the threshold is deduced from a user-selected probability of false alarm.

We have evaluated our method on classical synthetic sequences for which the performances of many detection methods are available [52], [56]. The comparative results on the dataset demonstrated that our method outperforms well-known unsupervised methods. We have also obtained very satisfactory results on real complex TIRF sequences.
Partners: Jean Salamero, J. Boulanger (UMR 144, PICT IBiSA, CNRS-Institut Curie)

6.3. Conditional random fields for vesicle traffic analysis with background estimation

Participants: Thierry Pécot, Patrick Bouthemy, Charles Kervrann.

Figure 6. Example of typical Time-Correlated Single Photon Counting (TCSPC) FLIM data. Total fluorescence intensity is shown in the center and corresponds to the sum of fluorescence intensities along the time axis at each pixel. The four side graphs correspond to time dependent photon counts in four different regions with variable sizes. By considering large regions, we observe an exponential decreasing along the time of fluorescence lifetime (see D). A: one pixel region; B and C: 3×3 *patches at different locations; D:* 15×15 *patch and liftetime estimation by least-square fitting.*

Figure 7. Segmentation method applied to a real TIRF microscopy sequence showing the Rab11-mCherry protein. The SLT-LoG method is able to provide the entire spatial support of the vesicles.

Figure 8. Left: Real fluorescence microscopy image depicting GFP-Rab6 proteins. Center: estimated vesicular component. Right: estimated background component.

Image analysis applied to fluorescence live cell microscopy has become a key tool in molecular biology since it enables to characterize biological processes in space and time at the subcellular level. In fluorescence microscopy imaging, the moving tagged structures of interest, such as vesicles, appear as bright spots over a static or non-static background. In this work, we consider the problem of vesicle segmentation and time-varying background estimation at the cellular scale. The main idea is to formulate the joint segmentation-estimation problem in the general Conditional Random Field (CRF) framework. Furthermore, segmentation of vesicles and background estimation are alternatively performed by energy minimization using a min cut-max flow algorithm. The proposed approach relies on a detection measure computed from intensity contrasts between neighboring patches in fluorescence microscopy images. We have demonstrated the competitiveness of the proposed method through an experimental comparison with state-of-the-art methods in fluorescence videomicroscopy, for single cell studies. We have also characterized the density of Rab6 transport carriers spatially dispersed at the cell periphery, for two different specific adhesion geometries.

Partners: Jean Salamero, J. Boulanger (UMR 144, PICT IBiSA, CNRS-Institut Curie)

6.4. Exemplar-based occlusion handling and sparse continuous aggregation for optical flow computation

Participants: Denis Fortun, Patrick Bouthemy, Charles Kervrann.

Handling large displacements, motion details and occlusions all together remains an open issue for reliable computation of optical flow in a video sequence. Our recently investigated aggregation paradigm is an attractive approach supplying motion candidates at every pixel in a first step, and combining them in a second step to determine the global optical flow field [16]. We experimentally demonstrate that simple and purely local parametric estimations combined with patch correspondences are sufficient to produce highly accurate motion candidates. Nevertheless, the performances are limited by the presence of large occlusion areas. Therefore we have proposed an exemplar-based occlusion handling scheme integrated in the two steps of the aggregation process. At the first stage, local motion candidates sets are extended at the detected occluded pixels with candidates from non-occluded pixels, and specific occlusions due to camera motion are handled by estimating the dominant motion in the image. Local occlusion cues are extracted from this first step. Then, we define a global energy function which cooperatively selects the best motion candidates for each point while recovering the occlusion areas and ensuring smoothness properties. Results on small displacement sequences are competitive with state-of-the-art methods, and great improvements are observed in the case of large displacements and occlusions (Fig. 9).

Alternatively to the discrete aggregation based on graph cut optimization, a new continuous aggregation model has been designed. In accordance with the demonstrated evidence that the set of candidates always contains at least one accurate motion vector, the aggregation is formulated in a sparse framework restricting the number of non negligible weights associated to the candidates. The continuous framework is less dependent on the quality of the candidates and thus allows us to considerably reduce the computational cost of both aggregation and candidates estimation.
Reference: [16]

6.5. Correlation and variational approaches for motion and diffusion estimation

Participants: Denis Fortun, Charles Kervrann.

Diffusion coefficient estimation in live cell fluorescence imaging is usually achieved with correlation-based methods related to Image Correlation Spectroscopy (ICS) [42]. This approach requires a high computational cost and the spatial resolution of the resulting diffusion map is limited by the inherent block-based principle of the method. To overcome these drawbacks, we propose a novel diffusion estimation method in a variational framework providing dense and discontinuity-preserving diffusion fields. The diffusion equation is integrated in a global energy via a neighborhood-wise data term, positivity constraint and temporal integration. The

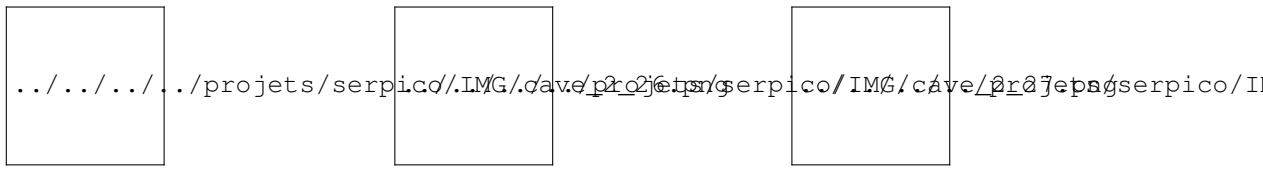
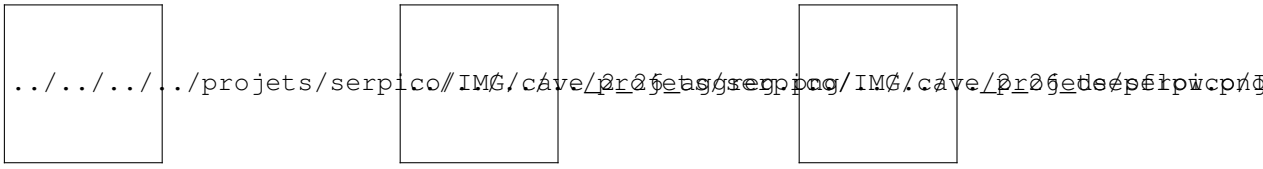
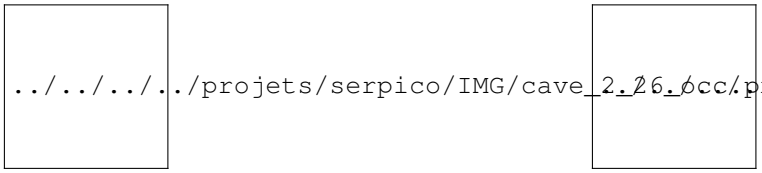


Image I_1 | Image I_2 | Optical flow ground-truth



Our method | Weinzaepfel et al. [63] | Xu et al. [64]



Occlusion ground truth | Occlusions recovered by our method

Figure 9. Comparative evaluation of optical flow estimation for large displacements between our method, [63] and [64]. First row : two successive frames I_1 and I_2 and the ground truth motion field; second row: comparative estimation results; third row: evaluation of our occlusion map estimation.

performances of the variational and ICS approaches were compared on simulated sequences. We have demonstrated the accuracy of ICS in stationarity conditions, and we pointed out the advantages of dense variational estimation to accurately recover spatial and temporal discontinuities (Fig. 10).
Reference: [17]
Partners: Perrine Paul-Gilloteaux, Francois Waharte and Chen Chen (UMR 144, PICT IBiSA, CNRS-Institut Curie)

6.6. Classification of membrane dynamics in TIRF microscopy

Participants: Antoine Basset, Charles Kervrann, Patrick Bouthemy.

Recognizing dynamic protein behaviors in live cell fluorescence microscopy is of paramount importance to understand cell mechanisms. In the case of membrane traffic, cargo molecules are transferred from a donor to an acceptor compartments [49]. At each step, dedicated molecular platforms are acting to form, transport and address selected proteins. In microscopy imaging, this sequence of processes leads to a series of heterogeneous dynamics, which need to be untangled in order to understand the spatiotemporal coordination of the molecular actors. In this study, we aim at locating and recognizing temporal events in TIRF microscopy image sequences related to membrane dynamics. After segmenting the time-varying vesicles in the image, we exploit space-time information extracted from three successive images only to model, locate and recognize the two dynamic configurations of interest: translational motion or local fluorescence diffusion (see Fig. 11). A likelihood ratio test is defined to solve this issue. Results on synthetic sequences and real TIRF sequences demonstrated the accuracy and efficiency of the proposed method.
Partners: Jean Salamero, J. Boulanger (UMR 144, PICT IBiSA, CNRS-Institut Curie)

6.7. Crowd motion classification

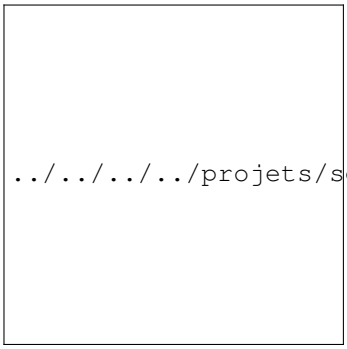
Participants: Antoine Basset, Charles Kervrann, Patrick Bouthemy.

Important research efforts have been devoted to crowd analysis for several years [58], [65]. We are interested in this topic for two main reasons. First, views of crowded scenes are not that different of light microscopy intracellular images. Second, the addressed problem, i.e. motion understanding, is common, and we are investigating similar data-driven methodological approaches. This a way to cross-fertilize two domains.

We address the problem of classifying coherent crowd motions in videos recorded by a fixed camera. In contrast to most existing methods, which are based on trajectories or tracklets, our approach for crowd motion analysis provides a crowd motion classification on a frame-by-frame basis. Indeed, we only compute affine motion models from pairs of two consecutive video images. The classification itself relies on simple rules on the coefficients of the computed affine motion models, and therefore does not imply any prior learning stage. The overall method proceeds in three steps: we first compute a set of motion model candidates on a collection of windows of different sizes in the image, then we select the motion model at each point owing to a ML criterion, finally we determine the crowd motion class map with a hierarchical classification tree regularized by majority votes. The algorithm is almost parameter-free, and is extremely efficient in terms of memory and computation load. Experiments on computer-generated sequences [28] and real video sequences demonstrate that our method is accurate, and can successfully handle complex situations (see Fig. 12).
References:[14], [23]

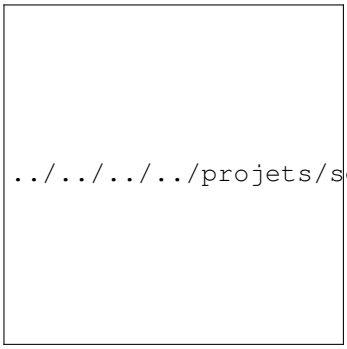
6.8. Estimation of the flow of particles without tracking algorithm in fluorescence imaging

Participants: Thierry Pécot, Patrick Bouthemy, Charles Kervrann.



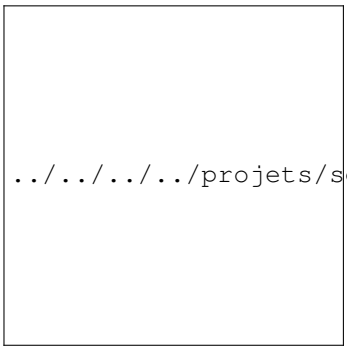
(a) Frame 1

(b) Ground-truth diffusion field



(c) Estimated diffusion field D

(d) Histogram of D



(e) Segmented diffusion field

(f) Profiles

Figure 10. Variational diffusion estimation on a simulated sequence with spatially variant diffusion. The curves of (f) are profiles of the dashed lines in (b),(c) and (e)

Figure 11. Classification results for a real TIRF sequence, whose estimated PSNR is 28.6. Results are displayed for a representative frame. The only classification error – framed in yellow – is a diffusion classified as translation. However, this vesicle has a very low intensity and changes its shape while diffusing. Two vesicles framed in green are detected as a single connected component. The vesicle framed in red corresponds to diffusing vesicles

Figure 12. Two frames of the Marathon bend sequence. People run from upper left to upper right, describing a U. The movement is quite constant in the whole sequence and so is the classification: in the left branch, people go South (magenta), then turn counterclockwise (red) until the end of the bend. Some Eastward translation (yellow) is sometimes found here because of the large radius of curvature. Finally the North translation is recovered (blue). The points in the upper right corner of the image are classified as translations to the West (purple), but the translation direction is closer to North than to West (North-North-West): it is also due to the lateral presence of pedestrians walking to the left.

Figure 13. Vesicle flows estimated with our method when considering a simple partition of 5 regions for an image sequence acquired in TIRF microscopy and showing the protein Clip170.

Automatic analysis of the dynamic content in fluorescence video-microscopy is crucial for understanding molecular mechanisms involved in cell functions. We have proposed an original approach for analyzing particle trafficking in these sequences. Instead of individually tracking every particle, we only need to locally count particles on regions over time and minimize a global energy function. We have specified three methods to determine the particle flow. We especially compared the NNLS algorithm [44] and the PPXA algorithm [33] known as well suited to non differentiable convex minimization problem [24]. We have conducted comparative experiments on synthetic and real fluorescence image sequences. We have shown that adding a sparsity constraint on the number of detected events allows us to reduce the number of false alarms. Compared to usual tracking methods, our approach is simpler and the results are very stable with respect to the only two parameters involved (see Fig. 13).
Reference: [24]

Partners: Jean Salamero, J. Boulanger (UMR 144, PICT IBiSA, CNRS-Institut Curie)

6.9. Probabilistic Tracking of fluorescent objects

Participants: Philippe Roudot, Charles Kervrann.

Image tracking of fluorescent objects, from labeled molecules to organelles and entire cells, is an essential task in the analysis of cellular functions. During the last decade, several algorithms have been tailored to cope with different types of cellular and subcellular motion down to Brownian single molecule behavior [8]. One of the remaining big challenges in this area of technology development has been the tracking of extremely heterogeneous movements of objects in crowded scenes. We tested several state-of-the-art algorithms [36], [40] to follow dense populations of diffusing particles, which suddenly change to directed motion. A frequent cellular scenario with this property is the jerky motion of vesicles and viruses switching between cytoplasmic diffusion and motor-mediated, fast displacements (see Fig. 14).

Figure 14. Vimentin motility seems to present a large proportion of confined Brownian motion and rare, sudden, motor-mediated transport. Colored tracks have been computed with an advanced U-track parametrization (Unit length filament of Vimentin Y117L mutant fused to GFP and transfected into vimentin null epithelial cell (cell line SW13). Image acquired with a spinning disk confocal microscospe with a 100x objective zoom 1.5 (Numerical Aperture 1.4, pixel size 0.10905μm/pixel).

Figure 15. A) Example of tracks simulation presenting a density of 3 spots/μm^2. B) Correct linking percentage wrt density and motion type switching probability. Our method outperforms U-track by 15% in the hardest case. C) True positive and false positive ratio on the same simulation with a density of 3 spots/μm^2, comparing our method with U-track, U-track with an on-line process noise estimator and an IMM algorithm with forward-backward initialization.

Figure 16. Correct linking and false positive percentage wrt speed switching probability.

These switches are particularly challenging to detect because they occur rarely. The presence of numerous detected objects in the expected range of particle displacements makes the tracking ambiguous and induces wrong associations. Lowering the ambiguity by reducing the search range, on the other hand, is not an option, as this would increase the rate of false negatives.

We first explored the existing methods in the literature to analyze their strenghts and weakness for tracking objects with heterogeneous motion and high density. Based on the conclusion we draw, we proposed a new method build on the U-track platform [40]. More specifically, we propose an interacting multiple state model that exploits recursive tracking in multiple rounds in forward and backward temporal directions. As a result, it achieves convergence of the instantaneous speed estimate time-point-by-time-point. This allows us to predict and recover abrupt transitions from freely or confined diffusive to directed motion. To address the issue of a particle that disappears as a neighboring particle appears in the same image and thus to better detect track termination, we also exploit this recursive tracking by proposing a locally adaptive on-line estimation of the search window radius for assignment (a.k.a. gating), while most of state-of-the-art algorithms propose only a global search window radius or weak per-track search radius estimations. We have shown on simulated data that our method outperforms state-of-the-art algorithms that model motion heterogeneity on different scenarios, e.g. heterogeneous motion type (see Figure 15) and speed heterogeneity (see Fig. 16), while keeping the computational cost of a deterministic method (10% overhead with respect to U-track).

Partners: Gaudenz Danuser (Harvard Medical School, Boston, USA)

6.10. Microtubules modeling for variational assimilation analysis

Participants: Pierre Allain, Charles Kervrann.

Microtubules (MT) are highly dynamic tubulin polymers that are involved in many cellular processes such as mitosis, intracellular cell organization and vesicular transport. Nevertheless, the modeling of cytoskeleton and MT dynamics based on physical properties is difficult to achieve. We proposed to model microtubules as rigid and growing cylinders alike (Nedelec and Foethk 2007) [45] but including Newtonian dynamics. Using the Euler-Bernoulli beam theory, we have proposed then to model the rigidity of microtubules on a physical basis using forces, mass and acceleration. In addition, we linked microtubules growth and shrinkage to the presence of molecules (e.g. GTP-tubulin) in the cytosol. The overall model enables linking cytosol to microtubules dynamics in a constant state space, thus allowing usage of data assimilation techniques (see Fig. 17).

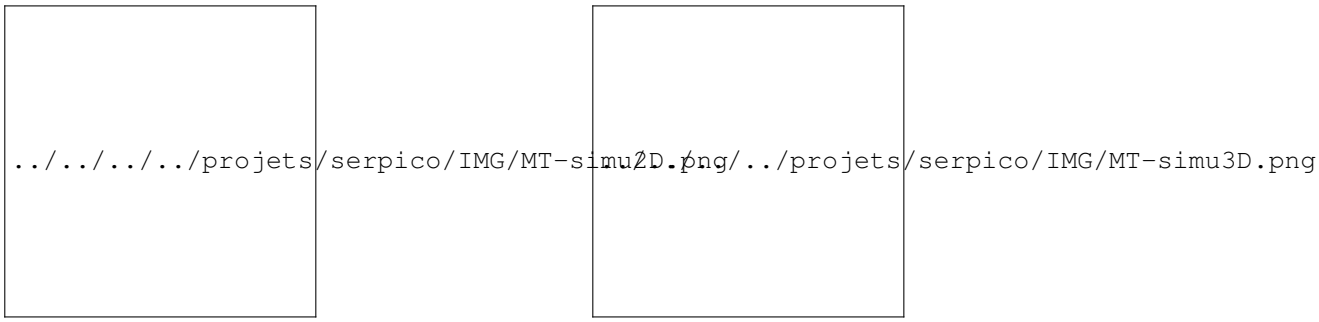


Figure 17. Left: Simulation of a 2D radial microtubule network. The results show growing and shrinking phases yielding inhomogeneous "pseudo-tubulin" concentration in the cytosol. MTs are bended according to fluid forces. Right: 3D simulation of MT nucleation and growth that mimics MT dynamics seeded onto a two vertical bar-shaped fibronectin pattern and observed in TIRF microscopy (courtesy of iRTSV/LPCV/PCM CEA-Grenoble).

6.11. Spot localization for TMA image analysis

Participants: Nam-Hoai Nguyen, Charles Kervrann.

A very first task of TMA (Tissue MicroArray) image analysis is to accurately localize spots (separate tissue core) representing arrays of 512 x 512 pixels each, in very large images of several thousands of pixels. For this purpose, we have investigated a three-stage methodological approach. First, since tissue cores are separately assembled in array (grid structure). We started to design a graphical model to eliminate image defects due to the presence of dusts or the imperfection of TMA blocks fabrication. In the second stage, a wavelet-like transform is currently used to recognize interested features (spots) given the size of spots *a prori*. Third, we started to investigate the superpixel-based image representation (SLIC) [27], [55] to handle very large images and biological details inside each spot.

Partners: V. Paveau (Innopys company)

SIROCCO Project-Team

6. New Results

6.1. Analysis and modeling for compact representation and navigation

3D modelling, multi-view plus depth videos, Layered depth images (LDI), 2D and 3D meshes, epitomes, image-based rendering, inpainting, view synthesis

6.1.1. Salient object detection

Participants: Olivier Le Meur, Zhi Liu.

Salient object detection consists in extracting in an automatic manner the most interesting object in an image or video sequence. From an input image, an object, with well-defined boundaries, is detected based on its saliency. This subject knows an renewed interest these last years. A number of datasets serving as ground truth has been released and can be used to benchmark methods.

In 2013, a new method to detect salient objects has been proposed [32], [18]. The principle relies upon low-level visual features and super-pixel segmentation. First, the original image is simplified by performing super-pixel segmentation and adaptive color quantization. On the basis of super-pixel representation, inter-super-pixel similarity measures are then calculated based on difference of histograms and spatial distance between each pair of super-pixels. For each super-pixel, its global contrast measure and spatial sparsity measure are evaluated, and refined with the integration of inter super-pixel similarity measures to finally generate the super-pixel-level saliency map. Experimental results on a dataset containing 1,000 test images with ground truths demonstrate that the proposed saliency model outperforms state-of-the-art saliency models. Figure 1 illustrates some results.

6.1.2. Image Memorability

Participant: Olivier Le Meur.

This work has been carried out in collaboration with Mattei Mancas (researcher of the University of Mons) during his visit of the team. The image memorability consists in the faculty of an image to be recalled after a period of time. Recently, the memorability of an image database was measured and some factors responsible for this memorability were highlighted. In [34] we proposed to improve an existing method by using attention-based visual features. To determine whether the visual attention plays a role in the memorability mechanism, eye tracking experiment has been performed by using a set of images of different memorability scores. Two important results have been observed. First the fixation duration is longer for the most memorable images (especially for the very first fixations) which shows a higher cognitive activity for memorable images. Second the observers congruency (agreement between observers) is significantly higher for the most memorable images. This shows that when there are areas with high attraction on all viewers, this induces higher memorability.

Following these first two observations, attention-based visual features were used to predict image memorability scores. A new set of features was then defined and used to train a model. Compared to an existing approach, we improve on the quality of the prediction of 2% while reducing the number of parameters by 14%. More specifically we replace the 512 features related to the GIST by 17 features which are directly related to visual attention.

6.1.3. Models for 3D video quality assessment

Participants: Darya Khaustova, Olivier Le Meur.

This work is carried out in collaboration with Orange labs. The goal is to design objective metrics for quality assessment of 3D video content, by establishing links between human visual perception (visual comfort) and video parameters such as quality and depth quantity, and between visual comfort and visual attention. The goal is also to study the differences in 2D visual attention in comparison with 3D visual attention.

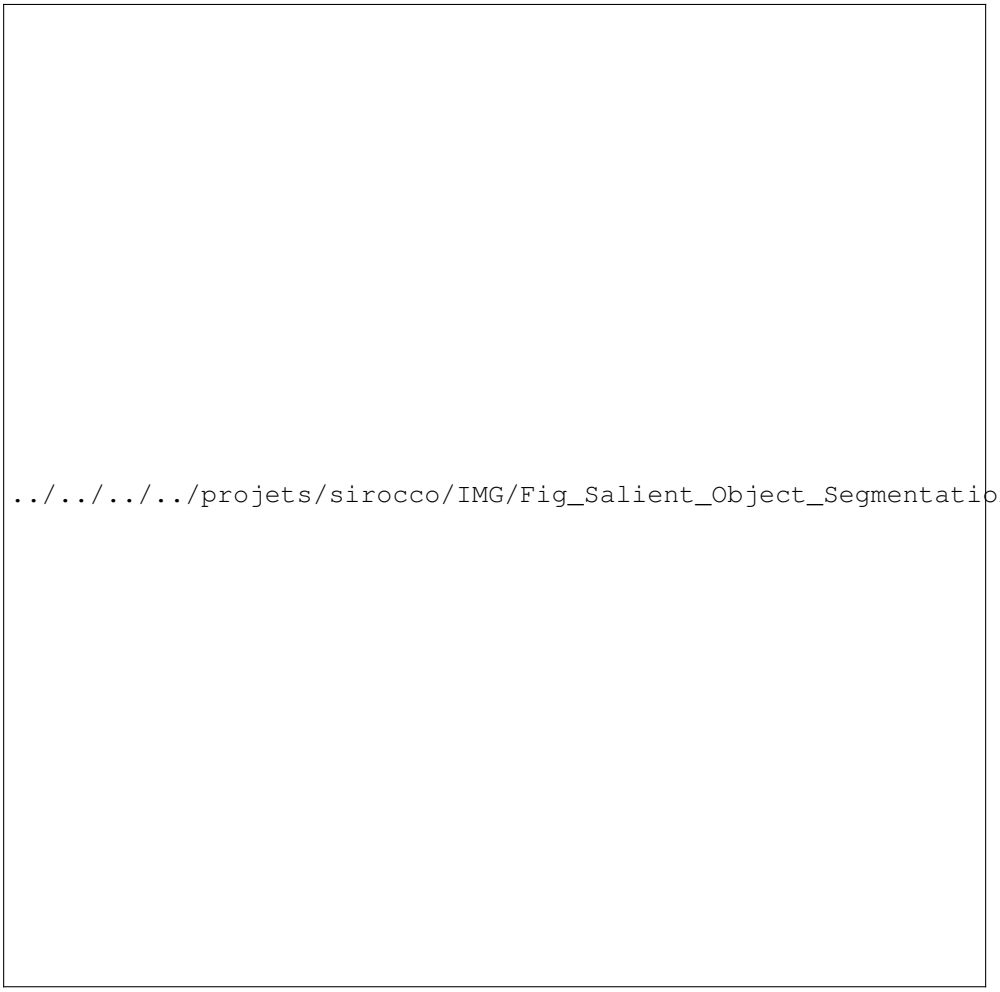


Figure 1. Illustration of the proposed approach: first row: original image; second row: saliency map; third row: extraction of the salient object.

Several subjective experiments have been carried out in order to study visual attention in different viewing conditions. The goal of the first experiment, involving 135 observers, was to study visual attention in three different conditions (2D, 3D comfortable and 3D uncomfortable), to eventually establish whether depth influences visual attention and whether there is a link between comfort and visual attention. The use of an eye-tracker allowed to record and to track observer's gaze. By analyzing the results, we found out that visual strategy to observe 2D images and 3D images with uncrossed disparity is very similar; there was no significant influence of discomfort on visual attention.

The second question which has then been addressed is how visual attention is influenced by objects with crossed disparity. A second test has been designed to answer this question, involving 51 observers. Considering scenes with crossed disparity it was revealed that objects located in front of the display plane are the most salient, even if observers experience discomfort. In the third experiment, we extended the study using scenes with crossed and uncrossed disparities. We verified the hypothesis that texture and contrast are more influential in guiding our gaze than the amount of depth. The features influencing the saliency of the objects in stereoscopic conditions were also evaluated with low-level visual stimuli. It was discovered that texture is the most salient feature in comparison to depth. Crossed disparity significantly influences the process of selecting the objects, while uncrossed disparity is less important, the process of selection being in this latter case similar to 2D conditions.

6.1.4. Epitome-based video representation

Participants: Martin Alain, Christine Guillemot.

This work is carried out in collaboration with Technicolor (D. Thoreau, Ph. Guillotel) and aims at studying novel spatio-temporal representations for videos based on epitomes. An epitome is a condensed representation of an image (or a video) signal containing the essence of the textural properties of this image. Different forms of epitomes have been proposed in the literature, such as a patch-based probability model learned either from still image patches or from space-time texture cubes taken from the input video. These probability models together with appropriate inference algorithms, are useful for content analysis inpainting or super-resolution. Another family of approaches makes use of computer vision techniques, like the KLT tracking algorithm, in order to recover self similarities within and across images. In parallel, another type of approach consists in extracting epitome-like signatures from images using sparse coding and dictionary learning.

We have in the past (in the context of the PhD thesis of S. Cherigui) developed a method for constructing epitomes for representing still images. The algorithm tracks self-similarities within the image using a block matching (BM) algorithm. The epitome is constructed from disjoint pieces of texture ("epitome charts") taken from the original image and a transform map which contains translational parameters (see Fig.2 . Those parameters keep track of the correspondences between each block of the input image and a block of the epitome. An Intra image compression scheme based on the epitome has been developed showing significant rate savings on some images, including the rate cost of the epitome texture and of the transform map. The entire image can be reconstructed from the epitome texture with the help of the transform map. The method is currently being extended to construct epitome representations of video segments rather than simple images. Such spatio-temporal epitome should pave the way for novel video coding architectures and open perspectives for other video processing problems which we have started to address such as denoising and super-resolution.

6.2. Rendering, inpainting and super-resolution

image-based rendering, inpainting, view synthesis, super-resolution

6.2.1. Image and video inpainting

Participants: Mounira Ebdelli, Christine Guillemot, Olivier Le Meur.

Figure 2. Original image and corresponding epitome.

Image (and video) inpainting refers to the process of restoring missing or damaged areas in an image (or a video). This field of research has been very active over the past years, boosted by numerous applications: restoring images from scratches or text overlays, loss concealment in a context of impaired image transmission, object removal in a context of editing, disocclusion in image-based rendering of viewpoints different from those captured by the cameras. Inpainting is an ill-posed inverse problem: given observations, or known samples in a spatial (or spatio-tempoal) neighborhood, the goal is to estimate unknown samples of the region to be filled in. Many methods already exist for image inpainting, either based on PDE (Partial Derivative Equation)-based diffusion schemes, either using sparse or low rank priors or following texture synthesis principles exploiting statistical or self-similarity priors.

Novel methods have been developed investigating two complementary directions first for image inpainting. The first direction which has been explored is the estimation of the unknown pixel with different neighbor embedding methods, i.e. Locally Linear embedding (LLE), LLE with a low-dimensional neigborhood representation (LLE-LDNR), Non-Negative Matrix Factorization (NMF) with various solvers [16]. The second method developed uses a two-steps hierarchical (or coarse to fine) approach to reduce the execution time [17]. In this hierarchical approach, a low resolution version of the input image is first inpainted, this first step being followed by a second one which recovers the high frequency details of the inpainted regions, using a single-image super-resolution method. To be less sensitive to the parameters setting of the inpainting, the low-resolution input picture is inpainted several times with different settings. Results are then efficiently combined with a loopy belief propagation. Experimental results in a context of image editing, texture synthesis and 3D view synthesis demonstrate the effectiveness of the proposed method.

The problem of video inpainting has also been considered. A first video inpainting algorithm has been developed in 2012, using a spatio-temporal examplar-based method. The algorithm proceeds in three steps. The first one inpaints missing pixels in moving objects using motion information. Then the static background is inpainted exploiting similarity between neighboring frames. The last step fills in the remaining holes in the current frame using spatial inpainting. This approach works well with static cameras but not so well when the video has been captured by free-moving cameras.

In 2013, we have therefore addressed the problem of video inpainting with free-moving cameras. The algorithm developed first compensates the camera motion between the current frame and its neighboring frames in a sliding window, using a new region-based homography computation which better respects the geometry of the scene compared to state-of-the-art methods. The source frame is first segmented into regions in order to find homogeneous regions. Then, the homography for mapping each region into the target frame is estimated. The overlapping of all aligned regions forms the registration of the source frame into the target one. Once the neighboring frames have been aligned, they form a stack of images from which the best candidate pixels are searched in order to replace the missing ones. The best candidate pixel is found by minimizing a cost function which combines two energy terms. One energy term, called the data term, captures how stationary is the background information after registration, hence enforcing temporal coherency. The second term aims at favoring spatial consistency and preventing incoherent seams, by computing the energy of the difference between each candidate pixel and its 4-neighboring pixels in the missing region. The minimization of the energy term is performed globally using Markov Random Fields and graph cuts. The proposed approach, although less complex than state-of-the-art methods, provides more natural results (see Fig.3).

Figure 3. Mask of the image to be inpainted; Results with the proposed video inpainting algorithm.

6.2.2. Image priors for inpainting

Participants: Raul Martinez Noriega, Aline Roumy.

Image inpainting is an ill-posed inverse problem which has no well-defined unique solution. To make this problem more "well-defined" it is necessay to introduce image priors. We consider here the problem of extracting such priors to help restoring the connection of long edges across the missing region. The prior is defined as a binary image that contains the locations of salient edge points located at the boundary of the missing region as well as the linear edges that join these points across the missing region. A method has been developed to extract such priors. It first detect edges which are then successively pruned in order to keep only informative edges, i.e., which have coherent gradients and are either part of a salient structure, or at the border between two different textures. Edges which are quasi-perpendicular to the boundary of the missing region are finally retained. Directions of the retained edges are computed and pairs of edges with similar directions are then connected with straight lines. These lines are used to segment the image into different regions and to define the processing order of the patches to be inpainted. Only patches from the known part and belonging

to the same region as the input patch are used. This avoids bringing details of one texture into another one, as well as the unconnected edge problem [35].

6.2.3. Image and video super-resolution

Participants: Marco Bevilacqua, Christine Guillemot, Aline Roumy.

Super-resolution (SR) refers to the problem of creating a high-resolution (HR) image, given one or multiple low-resolution (LR) images as input. The SR process aims at adding to the LR input(s) new plausible high-frequency details, to a greater extent than traditional interpolation methods (see, for example, Fig. 4 for a comparison between bicubic interpolation and SR). We mostly focused on the single-image problem, where only a single LR image is available.

We have adopted the example-based framework, where the relation between the LR and HR image spaces is modeled with the help of pairs of small "examples", i.e. texture patches. Each example pair consists of a LR patch and its HR version that also includes high-frequency details; the pairs of patches form a dictionary of patches. For each patch of the LR input image, one or several similar patches are found in the dictionary, by performing a nearest neighbor search. The corresponding HR patches in the dictionary are then combined to form a HR output patch; and finally all the reconstructed HR patches are re-assembled to build the super-resolved image.

In this procedure, one important aspect is how the dictionary of patches is built. At this regard, two choices are possible: an external dictionary, formed by sampling HR and LR patches from external training images; and an internal dictionary, where the LR/HR patch correspondences are learnt by putting in relation directly the input image and scaled versions of it. The advantage of having an external dictionary is that it is built in advance: this leads to a reduction of the computational time, whereas in the internal case the dictionary is generated online at each run of the algorithm. However, external dictionaries have a considerable drawback: they are fixed and so non-adapted to the input image. To be able to satisfactorily process any input image, we need then to include in the dictionary a large variety of patch correspondences, leading to a high computational time.

To overcome this problem, in [23] we proposed a novel method to build a compact external dictionary. The method consists in first jointly clustering LR and HR patches. The aim of this procedure, which we called JKC (Jointly K-means Clustering), is to prune the dictionary of the "bad" pairs of patches, i.e. those ones for which the cluster assignments of the related LR and HR patches do not correspond. Once the dictionary is clustered, it is summarized, by sampling some prototype patches, and applying on them simple geometrical transformations, in order to enrich the dictionary. The so constructed compact dictionary is shown to give equivalent or even better performance than the initial large dictionary with any input image.

The dictionary construction method described in [23] has been used as a basis for designing a full single-image SR algorithm. The new algorithm, presented in [25], follows the traditional scheme of example-based SR with an external dictionary, where a new way to generate the training patches is introduced. Given a HR training image H, the corresponding LR image L is generated; but instead of directly sampling patches from H and L, as usually done, the training images are further processed. An enhanced interpolation of L, using an iterated back projection, is used as a source of LR patches, and a high-frequency residual image, given by the difference between H and the interpolated LR image, is used for extracting HR patches. The JKC procedure is then applied to get the final compact dictionary. A special example-based SR algorithm has been designed, where the final HR output patches are constructed by combining selected HR residual patches from the dictionary with nonnegative weights. In the context of this study, we have also introduced a novel non-negative dictionary learning method [24]. The proposed method consists of two steps which are alternatively iterated: a sparse coding and a dictionary update stage. As for the dictionary update, an original method has been proposed, which we called K-WEB, as it involves the computation of k WEighted Barycenters.

Besides SR for still images, a preliminary work on video sequences has been also conducted [26]. In particular, we have considered the case of a LR video sequence with periodic high-resolution (HR) key frames. Given this scenario, a specific SR procedure has been designed to upscale each intermediate frame, by using the internal dictionary constructed from the two neighbor key frames.

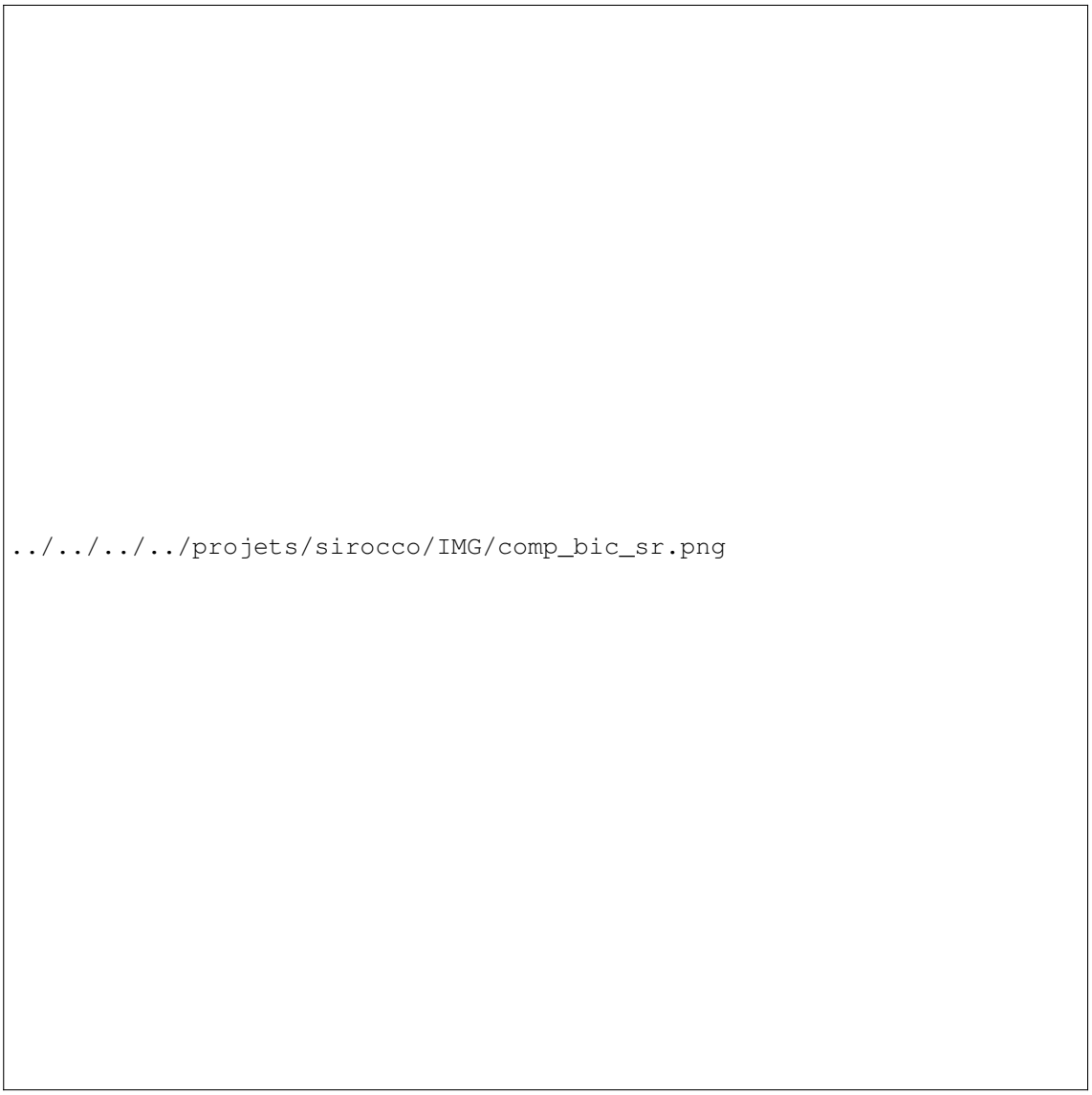


Figure 4. Comparison between bicubic interpolation and SR when upscaling the same LR image (by a factor of 3).

6.3. Representation and compression of large volumes of visual data

Sparse representations, data dimensionality reduction, compression, scalability, perceptual coding, rate-distortion theory

6.3.1. Multi-view plus depth video compression

Participants: Christine Guillemot, Laurent Guillo.

Multi-view plus depth video content represent very large volumes of input data wich need to be compressed for storage and tranmission to the rendering device. The huge amount of data contained in multi-view sequences indeed motivates the design of efficient representation and compression algorithms. The team has worked on motion vector prediction in the context of HEVC-compatible Multi-view plus depth (MVD) video compression. The HEVC compatible MVD compression solution implements a 6 candidate vector list for merge and skip modes. When a merge or a skip mode is selected, a merge index is written in the bitstream. This index is first binarized using a unary code, then encoded with the CABAC. A CABAC context is dedicated to the first bin of the unary coded index while the remaining bins are considered as equiprobable. This strategy is efficient as long as the candidate list is ordered by decreasing index occurrence probability. We have improved the construction of the candidate list by proposing two new candidates derived from disparity motion vectors in order to exploit inter-view correlation. This work has led to a joint proposal with Qualcomm and Mediatek which has been adopted in the HEVC-3DV standard in July 2013.

6.3.2. Spatio-temporal video prediction with neighbor embedding

Participants: Martin Alain, Christine Guillemot.

The problem of texture prediction can be regarded as a problem of texture synthesis. Given observations, or known samples in a spatial neighborhood, the goal is to estimate unknown samples of the block to be predicted. We have in 2012 developed texture prediction methods as well as inpainting algorithms using sparse representation as with learned dictionaries [19], or using neighbor embedding techniques [11], [30]. The methods which we have more particularly considered are Locally Linear Embedding (LLE), LLE with Low-dimensional neighborhood representation (LDNR), and Non-negative Matrix Factorization (NMF) using various solvers. In 2013, we have addressed the problem of temporal prediction for inter frame coding of video sequences using locally linear embedding (LLE). LLE-based prediction computes the predictor as a linear combination of K nearest neighbors (K-NN) searched within one or several reference frames. We have explored different K-NN search strategies in the context of temporal prediction, leading to several temporal predictor variants using or not motion information [22]. A parallel was also drawn between such multi-patch based prediction and the adaptive interpolation filtering (AIF) method. The LLE-based inter prediction techniques, when used as extra modes for inter prediction in an H.264 codec, are shown to bring significant Rate-Distortion (RD) performance gains compared to H.264 (up to 21.76 $\%$ bit-rate saving) and with respect to the use of AIF.

6.3.3. Dictionary learning for sparse coding of satellite images

Participants: Jeremy Aghaei Mazaheri, Christine Guillemot, Claude Labit.

In the context of the national partnership Inria-Astrium, we explore novel methods to encode images captured by a geostationary satellite. These pictures have to be compressed on-board before being sent to earth. Each picture has a high resolution, therefore the rate without compression is very high (about 70 Gbits/sec). The goal is to achieve a rate after compression of 600 Mbits/sec, i.e., a compression ratio higher than 100. On earth, the pictures are decompressed with a high reconstruction quality and visualized by photo-interpreters. The goal of the study is to design novel transforms based on sparse representations and learned dictionnaries for satellite images.

Sparse representation of a signal consists in representing a signal $y \in \Re^n$ as a linear combination of columns, known as atoms, from a dictionary matrix. The dictionary $D \in \Re^{n \times K}$ is generally overcomplete and contains K atoms. The approximation of the signal can thus be written $y \approx Dx$ and is sparse because a small number of atoms of D are used in the representation, meaning that the vector x has only a few non-zero coefficients. Sparsity of the representation depends on how the dictionary is representative of the data at hand, hence the need to learn appropriate dictionaries.

We have developed methods for learning adaptive tree-structured dictionaries, called Tree K-SVD [20]. Each dictionary in the structure is learned on a subset of residuals from the previous level, with the K-SVD algorithm. The tree structure offers better rate-distortion performance than a "flat" dictionary learned with K-SVD, especially when only a few atoms are selected among the first levels of the tree. The tree-structured dictionary allows efficient coding of the indices of the selected atoms. We recently developped a new sparse coding method adapted to this tree-structure to improve the results [20]. The tree-structured dictionary has been further improved by studying different branch pruning strategies. The use of these dictionaries in an HEVC-based intra coder is under study. The dictionaries are also considered for scene classification and for detecting the MTF (Modulation Transfer Function) of the optical capturing system.

6.3.4. HDR video compression

Participants: Christine Guillemot, Mikael Le Pendu.

High Dynamic Range (HDR) images contain more intensity levels than traditional image formats. Instead of 8 or 10 bit integers, floating point values are generally used to represent the pixel data. Floating point video formats are widely used in the visual effects industry. Moreover, the development of a new standardized workflow ACES intends to generalize the use of such formats to the whole cinema production pipeline. The increasing use of floating point representations, however, comes with a technical issue concerning the storage space required for those videos with higher precision than the current 8 or 10 bit standards.

In collaboration with Technicolor (D. Thoreau), we worked on floating point video compression. Different approaches exist in the literature. Several methods consists in compressing directly the floating point data using its internal representation (i.e. sign, exponent and mantissa bits). These methods are generally limited to lossless compression schemes. Another type of approach makes use of the existing compression standards such as H264/AVC or HEVC to encode a floating point sequence of images previously converted to lower bit depth integers. In this approach, the conversion is designed to be reversible with minimal loss. However the converted integer images are not intended for being displayed directly. Finally a last family of approach aims at keeping backward compatibility with an existing compression standard. The original image sequence is first tone mapped and encoded to obtain a low dynamic range (LDR) version that can be visualized on a standard LDR display. In parallel, a residual information needed to reconstruct the HDR image from the LDR version is also encoded.

In our study, a floating point to integer conversion method was developed to be applied before HEVC compression. The original floating point RGB values are converted to high bit depth integers with an approximate logarithmic encoding that is reversible without loss. The RGB values are then converted to a YUV color space. The bit depth must also be reduced to be supported by the compression standard. This bit depth reduction is performed adaptively depending on the minimum and maximum values (i.e. darkest and brightest points respectively) which characterize the real dynamic of the data. In the best case, the difference between the extreme values is sufficiently low to perform this operation without loss.

Three variants of the method have been compared. The conversion can be performed either by Groups of Pictures (GOP), or independently on each frame of the sequence, or even more locally, by blocks of pixels. The GOP-wise approach combined with spatial and temporal predictions in the encoder gives the best results for low bit rate compression. The block-wise approach can reduce the bit depth with less data loss but breaks the continuity between the blocks, which degrades the Rate Distortion (RD) performance especially at low bit rates. However, we have shown that this approach gives the best results in the context of near lossless compression. The frame-wise version is intermediate between the global (GOP-wise) and local (block-wise) versions. It is adapted to high quality compression. This method was also compared to another frame-wise

conversion method in the recent literature called adaptive LogLuv transform, and a 50% rate saving was obtained at high bitrates.

6.3.5. HEVC coding optimization

Participants: Nicolas Dhollande, Christine Guillemot, Bihong Huang, Olivier Le Meur.

The team has two collaborations in the area of HEVC-based video coding optimization. The first research activity is carried out in collaboration with Orange labs (Felix Henry) and UPC (Philippe Salembier) in Barcelona. The objective is to design novel methods for predicting the residues resulting from spatio-temporal prediction. We have indeed observed that the redundancy in residual signals (hence the potential rate saving) is high. In 2013, different methods have been investigated to remove this redundancy, such as generalized lifting and different types of predictors. The generalized lifting is an extension of the lifting scheme of classical wavelet transforms which permits the creation of nonlinear and signal probability density function (pdf) dependent and adaptive transforms.

The second collaboration is with Thomson Video Networks and aims at designing an innovative architecture for effective real-time broadcast encoders of Ultra High Definition (UHD) contents. Currently, the only way to transmit acceptable UHD contents around $10-20$ Mbits/sec is the new compression standard HEVC (finalized in January 2013). Yet, UHD requires at minimum 8 times more computation than the actual HDTV formats, and HEVC has a computing complexity which is already from 2 to 10 times that of MPEG4-AVC. To reduce the encoding complexity on UHD content, a pre-analysis with a lower resolution version (HD) of the input content has been considered to infer some decisions and coding parameters on the UHD video. A speed-up of a factor 3 has already been achieved for a small rate loss of $4-5\%$.

6.4. Distributed processing and robust communication

Information theory, stochastic modelling, robust detection, maximum likelihood estimation, generalized likelihood ratio test, error and erasure resilient coding and decoding, multiple description coding, Slepian-Wolf coding, Wyner-Ziv coding, information theory, MAC channels

6.4.1. Loss concealment based on video inpainting

Participants: Mounira Ebdelli, Christine Guillemot, Ronan Le Boulch, Olivier Le Meur.

We have developed a loss concealment scheme based on a new hierarchical video examplar-based inpainting algorithm. The problem of loss concealment is to estimate unknown pixels after decoding when the corresponding transport packets have been lost on the transmission network. Before proceeding to the video texture inpainting, the motion vectors of the lost blocks must first be estimated from the motion vectors of the received blocks in the spatial neighborhood. The Motion vectors (MV) of damaged blocks are estimated using a Bilinear Motion Field Interpolation (BMFI) technique.

The algorithm follows a coarse to fine approach and first inpaints a low resolution version of the damaged video. Moving objects, detected thanks to the estimated motion vectors, are processed first. The most similar patches (similar to the known pixels of the patch to be completed) is searched within a motion-compensated window in adjacent frames, and used as an estimate of the pixels to be filled in. Then the static background is inpainted using known co-located pixels of neighboring frames. The remaining holes are filled-in using spatial inpainting.

In a second step, the high frequency details of the inpainted areas are recovered using a super-resolution technique, in the same vein as described in Section 6.2.1 for still images. The inpainted low resolution video is first interpolated using a simple lanczos interpolation. The idea is then to search for the nearest neighbor (the best match) of the interpolated version of each inpainted block, within the known part of the current image of the impaired video at the native resolution. The found correspondences form a so-called nearest neighbor field (NNF) which connects inpainted and interpolated patches of the low resolution video to high resolution patches of known parts of the high resolution (HR) video. The found NN patch is then copied to replace the low resolution inpainted patch.The two-step approach allows significantly reducing the execution time of the video inpainting process, while preserving a satisfactory quality.

6.4.2. Universal distributed coding

Participant: Aline Roumy.

In 2012, we started a new collaboration with Michel Kieffer and Elsa Dupraz (Supelec, L2S) on universal distributed source coding. Distributed source coding refers to the problem where several correlated sources need to be compressed without any cooperation at the encoders. Decoding is however performed jointly. This problem arises in sensor networks but also in video compression techniques, where the correlation between the successive frames is not directly used at the encoder, and are therefore seen as distributed. Traditional approaches (from an information theoretical but also practical point of view) assume that the correlation channel between the sources is perfectly known. Since this assumption is not satisfied in practice, a way to get around this is to use a feedback channel (from the decoder to the encoder), that can trigger the encoder.

Instead, we consider universal distributed source coding, where the correlation channel is unknown and belongs to a class parametrized by some unknown parameter vector. We proposed four uncertainty models that depend on the partial knowledge we have on the correlation channel and derived the information theoretical bounds [28]. A complete coding scheme has also been proposed that works well for any distribution in the class [27]. At the encoder, the proposed scheme encompasses the determination of the coding rate and the design of the encoding process. Both contributions result from the information-theoretical compression bounds of universal lossless source coding with side information. Then a novel decoder is proposed that takes into account the available information regarding the class. The proposed scheme avoids the use of a feedback channel or the transmission of a learning sequence, which both would result in a rate increase at finite length.

SUMO Team

6. New Results

6.1. Model expressivity and quantitative verification

6.1.1. Diagnosis from scenarios

Participants: Loïc Hélouët, Blaise Genest, Hervé Marchand.

Diagnosis of a system consists in providing explanations to a supervisor from a partial observation of the system and a model of possible executions. This year, we have extended results on diagnosis algorithm from scenarios. Systems are modeled using High-level Message Sequence Charts (HMSCs), and the diagnosis is given as a new HMSC, which behaviors are all explanations of the partial observation. The results published this year are first an offline centralized diagnosis algorithm (a single process in a network collects an observation, and emits a diagnosis) that has then been extended to a decentralized version of this algorithm. This allows us to give a complete diagnosis framework for infinite state systems, with a strong emphasis on concurrency and causal ordering in behaviors. HMSC-based diagnosis showed nice properties w.r.t. compositionality. We have also considered solutions for online diagnosis from scenarios, but came to the conclusion that online solutions are memory consuming, and need too many restrictions to run with finite memory.

The last contribution of this work is an application of diagnosis techniques to anomaly detection, that is a comparison of observation of the system with a model of usual behaviors to detect security attacks. This work is already available online in [25], and will soon be published.

6.1.2. Probabilistic model checking

Participants: Nathalie Bertrand, Blaise Genest, Paulin Fournier.

In [20], we considered the verification of Markov chains against properties talking about distributions of probabilities. Even though a Markov chain is a very simple formalism, by discretizing in a finite number of classes the space of distributions through some symbolics, we proved that the language of trajectories of distribution (one for each initial distribution) is not regular in general, even with 3 states. We then proposed a parametrized algorithm which approximate what happens to infinity, such that each symbolic block in the approximate language is at most ϵ away from the concrete distribution.

With the objective of model checking infinite state probabilistic systems, we proved a general finite-time convergence theorem for fixpoint expressions over a well-quasi-ordered set [22]. This has immediate applications for the verification of well-structured systems, where a main issue is the computability of fixpoint expressions, and in particular for game-theoretical properties and probabilistic systems where nesting and alternation of least and greatest fixpoints are common [35].

Parameterized verification aims at validating a system's model irrespective of the value of a parameter. In [34] we introduced a model for networks of an arbitrary number of probabilistic timed processes, communicating by broadcasting. This model is suitable for distributed protocols, and can be applied to wireless sensor networks or peer-to-peer applications. The number of processes is unknown and either is constant (static case), or evolves over time through random disappearances and creations (dynamic case). On the one hand, most parameterized verification problems turn out to be undecidable in the static case (even for untimed processes). On the other hand, we prove their decidability in the dynamic case.

6.1.3. Distributed timed systems

Participants: Nathalie Bertrand, Amélie Stainer.

We study the reachability problem for communicating timed processes, both in discrete and dense time. Our model comprises automata with local timing constraints communicating over unbounded FIFO channels. Each automaton can only access its set of local clocks; all clocks evolve at the same rate. Our main contribution is a complete characterization of decidable and undecidable communication topologies, for both discrete and dense time. We also obtain complexity results, by showing that communicating timed processes are at least as hard as Petri nets; in the discrete time, we also show equivalence with Petri nets. Our results follow from mutual topology-preserving reductions between timed automata and (untimed) counter automata. To account for urgency of receptions, we also investigate the case where processes can test emptiness of channels. This resut is published in [39] and is a part of Amélie Stainer's PhD manuscript [18]. It also constitutes a contribution to ANR VACSIM.

We also studied a model for distributed systems composed of stochastic and timed processes that interact via broadcasting. For these networks of stochastic timed automata (NSTA), we provided a precise performance evaluation algorithm, without resorting to simulation techniques. The idea is to characterize the general state space Markov chain through transient stochastic state classes that represent the system's state after each action. This yields an algorithmic approach to the transient analysis of NSTA models, with fairly general termination conditions [32].

6.2. Management of large distributed systems

6.2.1. Test generation from Recursive Tile Systems

Participants: Sébastien Chédor, Thierry Jéron, Christophe Morvan.

We explore the generation of conformance test cases for Recursive Tile Systems (RTSs) in the framework of the classical ioco testing theory. The RTS model allows the description of reactive systems with recursion, and is very similar to other models like Pushdown Automata, Hyperedge Replacement Grammars or Recursive State Machines. Test generation for this kind of models is seldom explored in the literature. We first propose an off-line test generation algorithm for Weighted RTSs, a determinizable sub-class of RTSs, and second, an on-line test generation algorithm for the full RTS model. Both algorithms use test purposes to guide test selection through targeted behaviours. Additionally, essential properties relating verdicts produced by generated test cases on an implementation with both the conformance with respect to its specification, and the precision with respect to a test purpose, are proved. This work is published in [51], and a journal version will appear in 2014. It is also a part of Sébastien Chédor's PhD manuscript.

6.2.2. Distributed control

Participants: Blaise Genest, Hervé Marchand.

We focused this year on the control of distributed systems modeled as *asynchronous automata*, that is asynchronous network of automata communicating through peer to peer synchronizations. First, we considered the case where all events are controllable, and the objective is to accept exactly a given language. Here, a famous result is the Zielonka theorem [62], stating that every regular language closed under commutation can be turned into an asynchronous automaton. However, the construction is plagued with deadends and final state of the network are decided by a global controller monitoring every process at the same time and perfectly, which is unrealistic and defeat the distribution idea. This year, we characterized the languages which can be controlled realistically (no deadends, local final states and local decision on each process), and give algorithms to obtain the associated distributed machines in [30]. The case where some events are uncontrollable is reputed very difficult. We made a progress this year in [42], showing that we can decide whether a reachability objective can be ensured, granted that the communication between the processes follow a tree: siblings can not communicate directly together, they need to go through their common parent.

In [27], we consider an alternative model for the control of distributed systems; the aim is to build local controllers that restrict the behavior of a distributed system in order to satisfy a global state avoidance property. We model distributed systems as communicating finite state machines with reliable unbounded FIFO queues between subsystems. Local controllers can only observe the behavior of their proper subsystem and do not see the queue contents. To refine their control policy, controllers can use the FIFO queues to communicate by piggy-backing extra information (some timestamps and their state estimates) to the messages sent by the subsystems. We provide an algorithm that computes, for each local subsystem (and thus for each controller), during the execution of the system, an estimate of the current global state of the distributed system. We then define a synthesis algorithm to compute local controllers. Our method relies on the computation of (co-)reachable states. Since the reachability problem is undecidable in our model, we use abstract interpretation techniques to obtain overapproximations of (co-)reachable states. Similarly, in [46], we have been interested in the control of distributed systems with synchronous communications (called decentralized Discrete Event Systems). We introduced a novel architecture that extends the class of problems that can be solved in decentralized DES control in the absence of communication. In this architecture, unlike previous architectures that use either conjunction or disjunction to fuse local control decisions, the fusion rule is exclusive or. We characterized the new architecture, where controllers take a single decision, with respect to the recently-proposed multi-decision framework of Chakib and Khoumsi. Unlike previous architectures, parity-based controllers cannot predetermine their local control decision based solely on their local observations. Instead, the local control decisions are calculated a priori.

6.2.3. *Enforcement of timed and security properties*

Participants: Thierry Jéron, Hervé Marchand, Srinivas Pinisetty.

Runtime enforcement is a verification/validation technique aiming at correcting (possibly incorrect) executions of a system of interest. This year, we first consider enforcement monitoring for systems with timing specifications (modeled as timed automata). We consider runtime enforcement of any regular timed property specified by a timed automaton [45]. To ease their design and their correctness-proof, enforcement mechanisms are described at several levels: enforcement functions that specify the input-output behavior, constraints that should be satisfied by such functions, enforcement monitors that implement an enforcement function as a transition system, and enforcement algorithms that describe the implementation of enforcement monitors. The feasibility of enforcement monitoring for timed properties is validated by prototyping the synthesis of enforcement monitors. This work is also a contribution to ANR Vacsim. In [41], we studied an alternative enforcement problem of security properties, namely, the enforcement of K-step opacity at runtime. In K-step opacity, the knowledge of the secret is of interest to the attacker within K steps after the secret occurs and becomes obsolete afterwards. We introduce the mechanism of runtime enforcer that is placed between the output of the system and the attacker and enforces opacity using delays. If an output event from the system violates K-step opacity, the enforcer stores the event in the memory, for the minimal number of system steps until the secret is no longer interesting to the attacker (or, K-step opacity holds again)

6.2.4. *Discrete control of computing systems administration*

Participants: Hervé Marchand, Nicolas Berthier.

We address the problem of using Discrete Controller Synthesis for the administration of Computing Systems, following an approach supported by a programming language [24]. We present a mixed imperative/declarative programming language, where declarative contracts are enforced upon imperatively described behaviors. Its compilation is based on the notion of supervisory control of discrete event systems. More precisely, our language can serve programming closed-loop adaptation controllers, enabling flexible execution of functionalities w.r.t. changing resource and environment conditions. DCS is integrated into a1 programming language compiler, which facilitates its use by users and programmers, performing executable code generation. The tool is concretely built upon the basis of a reactive programming language compiler, where the nodes describe behaviors that can be modeled in terms of transition systems. Our compiler integrates this with a DCS tool, making it a new environment for formal methods. We apply our method to the problem of coordinating several administration loops in a data center (number of servers, repair, and local processor frequencies) [40].

We formulate this problem as an invariance controller synthesis problem. We are currently working on an extension of the controller synthesis tool so that it can handle the use of numerical variables in order to model both the system and the properties to be ensured by control.

6.2.5. Distributed planning

Participant: Éric Fabre.

Planning problems consist in organizing actions in a system in order to reach one of some target states. The actions consume and produce resources, can of course take place concurrently, and may have costs. We have a collection of results addressing this problem in the setting of distributed systems. This takes the shape of a network of components, each one holding private actions operating over its own resources, and shared/synchronized actions that can only occur in agreement with its neighbors. The goal is to design in a distributed manner a tuple of consistent local plans, one per component, such that their combination forms a global plan of minimal cost.

Our previous solutions to this problem modeled components as weighted automata. In collaboration with Loïg Jezequel (TU Munich) and Victor Khomenko (Univ. of Newcastle), we have extended this approach to the case of components modeled as safe Petri nets [44]. This allows one to benefit from the internal concurrency of actions within a component. Benchmarks have shown that this method can lead to significant time reductions to find feasible plans, in good cases. In the least favorable cases, performances are comparable to those obtained with components modeled as automata. The method does not apply to all situations however, as computations require to perform ϵ-reductions on Petri nets.

6.2.6. Diagnosis based on self-modeling

Participants: Éric Fabre, Carole Hounkonnou.

Model-based approaches have been proved to provide the best results for fault diagnosis in telecommunication networks, with various kinds of models. They suffer however from several difficulties: one has to build a model adequate to the supervised network (and possibly adapt it as the network evolves), one has to find the correct abstraction level for this model, and one has to deal with size issues of such models. In Carole Hounkonnou's thesis [15], we have proposed an approach that addresses these three limitations, under the generic name of self-modeling. It consists modeling a network in a generic manner, through its building rules. The actual instance one has to manage is then discovered on the fly, when some malfunction explanation request is triggered. Starting from the identified malfunction, the network model instance is discovered/revealed progressively, as requested by the needs of the diagnosis procedure. The latter progressively extends a Bayesian network model of the network, in order to collect more information and identify the malfunction rootcause. The model extension is guided by an information theory criterion: it seeks access to the new observations that are be the most informative (on the average) given previous observations taken into account. This approach allows to deal with potentially large models, as the supervised system needs not be entirely modeled before the diagnosis starts. We are currently working on the extension of this setting to model refinement, and to a framework of dynamic systems rather than static systems.

6.2.7. Graceful restart methods for link state routing protocols

Participants: Éric Fabre, Carole Hounkonnou.

Link state routing protocols are ubiquitous in the internet. OSPF (Open Shortest Path First) is one of them within an Autonomous System. In collaboration with Alcatel-Lucent, we have proposed an extension of graceful restart procedures, that allow to shut down the control plane of routers while maintaining the data plane active, and thus the packet forwarding activity. A drawback of existing procedures was that frozen routers had to be removed from the network as soon as topology evolved. We have shown that this pessimistic precaution could be damageable to the network and was not necessary [43]. Frozen routers may still be useful, even if they do not forward packets in an optimal manner. And even if they create routing loops, the latter can be easily detected, and optimally patched, which is often more efficient than declaring these routers as dead. Experiments on classical topologies of the topology zoo, as well as on random topologies, have confirmed these results.

6.3. Data driven systems

6.3.1. Web services

Participants: Blaise Genest, Loïc Hélouët.

This year, we considered transactional properties (ACID) for web services. In particular, we focused on the atomicity (A of ACID) property, obtained in case of a failure inside an atomic block through compensation of the executed actions of the block. To do so, logs need to be kept. We were interested in maintaining the maximal amount of privacy. We proposed modular algorithms [23] which maintain privacy between modules, with minimal information shared among modules, both in the logging and the compensation phases. Furthermore, each module logs a small number of information, such that the sum of all actions logged is guaranteed minimal. Last, modularity allows fast algorithms, as they need to consider only what happens in the module itself, and not the exact structure of its parent module nor of its sub-modules.

We also have extended the *session system* model originally proposed in [55]. We have deisgned a mode for Web-based systems that allows to describe systems running an arbitrary number of transactions over an arbitrary number of agents. For these systems, syntactic restrictions allow to decide coverability properties, and then more elaborated business rules, such as conflict of interest (the fact that a participant to a system can be involved in two exclusive services), or the Chinese Wall Property (that prevents users of a system to use benefits or information right they may have obtained from a privileged role at later instant of any execution of the system. These results were obtained with M. Mukund and S. Akshay within the context of the DISTOL associated team, and should lead to a publication next year.

6.3.2. Implementation of scenarios

Participants: Loïc Hélouët, Rouwaida Abdallah.

We have revisited the problem of program synthesis from specifications described by High-level Message Sequence Charts. The main objective is to obtain a distributed implementation (for instance described with communicating automata) from a global specification given as High-level MSCS. In the general case, synthesis by a simple projection on each component of the system allows more behaviors in the implementation than in the specification. The differences arise from loss of ordering among messages, but we have shown that for a subclass of HMSCs (the *local HMSCs*) behaviors can be preserved by addition of communication controllers, that intercept messages to add stamping information before resending them, and deliver messages to processes in the order described by the specification. This work was published in [19].

The second aspect of our work on scenarios implementability has considered implementation of requirements expressed as non-local HMSCs. We have proposed a new technique to transform an arbitrary HMSC specification into a local HMSC, hence allowing implementation. This transformation can be automated as a constraint optimization problem, and the impact of modifications brought to the original specification minimized w.r.t. a cost function. The approach was evaluated on a large number of randomly generated HMSCs, and the results show an average runtime of a few seconds, which demonstrates applicability of the technique. These results were published in [28]. Both results mentionned in this sections are part of the PhD thesis of Rouwaida Abdallah, defended this year [14].

6.3.3. Attribute grammars

Participant: Éric Badouel.

Evaluation of attributes w.r.t. an attribute grammar can be obtained by inductively computing a function expressing the dependencies of the synthesized attributes on inherited attributes. This higher-order functional approach to attribute evaluation can straightforwardly be implemented in a higher-order lazy functional language like Haskell. The resulting evaluation functions are, however, not easily amenable to optimization when we want to compose two attribute grammars. In [21], we present an alternative first-order functional interpretation of attribute grammars where the input tree is replaced by an extended cyclic tree each node of which is aware of its context viewed as an additional child tree. These cyclic representations of zippers (trees with their context) are natural generalizations of doubly-linked lists to trees over an arbitrary signature.

Then we show that, up to that representation, descriptional composition of attribute grammars reduces to the composition of tree transducers.

TASC Project-Team

6. New Results

6.1. Solvers

Participants: Nicolas Beldiceanu, Rémi Douence, Narendra Jussien, Xavier Lorca, Eric Monfroy, Charles Prud'Homme.

- [14] presents some research directions wrt sustainable solver development based on the idea that solvers should be based/derived on data bases of combinatorial knowledge.
- [19] and [42] presents a solver independent language dealing both with variable-oriented and constraint-oriented propagation engines to enable the design of propagation engines.
- By observing the resolution process, [35] shows how to dynamically adapt the resolution while propagating constraints.

6.2. Filtering

Participants: Nicolas Beldiceanu, Alban Derrien, Jérémie Du Boisberranger, Jean-Guillaume Fages, Arnaud Letort, Xavier Lorca, Thierry Petit, Charlotte Truchet, Mohamed Wahbi.

- Given a matrix model, with the same constraint defined by a finite-state automaton on each row and a global cardinality constraint on each column, [12] exploits double counting to derive necessary conditions on the cardinality variables of the global cardinality constraints from the automata. (participants: Beldiceanu)
- By using the observation that most global constraints can be reformulated as a conjunction of a total function constraint together with a constraint that can be easily reified (e.g. a linear constraint involving two variables), [13] introduces a simple way for deriving reified global constraints. (participants: Beldiceanu)
- In the context of distributed constraint solving [22], [25] introduce two filtering algorithms that extend Asynchronous Forward Checking (AFC). The last one outperforms AFC specially on sparse problems. (participants: Wahbi)
- We improve the energetic reasoning checker of the cumulative constraint by decreasing the number checked intervals by a factor seven. We prove this approach can be generalized to the ER filtering algorithm. Furthermore, in a context of makespan minimization of hard problems, our experiments demonstrate that associating this checker with a Time-Table propagator is more efficient than using the best state-of-the-art propagators, such as Time-Table Edge-Finding. This work is at the core of Alban Derrien's doctoral research (Alban is a PhD student of TASC). It was published at the doctoral program of CP2013 [28]. (participants: Derrien, Petit)
- [29] introduces a probabilistic model for the bound consistency algorithm of the alldifferent constraint in order to decrease the number of times the constraint is woken without making new deductions during constraint propagation. (participants: Du Boisberranger, Lorca, Truchet).
- Initially motivated by the shift minimisation personal task scheduling problem [30] shows how to integrate difference constraints into the AtMostNValue constraint in order to get a better estimation about the minimum number of distinct values. (participants: Fages, Lorca)
- Motivated by scalability issues, and based on the idea of accelerating the convergence to the fix-point by filtering several cumulative constraints in parallel, [33] and [47] presents a sweep based algorithm for a conjunction of cumulative constraints. (participants: Beldiceanu, Letort)
- [46] come up with a more efficient filtering algorithm than the one introduced for the cost regular constraint for dealing with constraints for which the set of solutions can be represented by an automaton with counters. (participants: Beldiceanu)

6.3. Continuous/discrete

Participants: Nicolas Beldiceanu, Gilles Chabert, Jean-Guillaume Fages, Charles Prud'Homme.

- While convexity (and some of its generalisations) is a key property used for dealing with continuous constraints it was not yet used in the context of discrete global constraints. In [34] we come up with a parametric filtering algorithm based on a form of convexity. It can handle in a uniform way various constraints such as deviation, spread or the conjunction of a linear inequality constraint and count constraint.
- Motivated by hybrid discrete continuous problems we come up in [44] with a simple and efficient interface for connecting a discrete constraint solver (Choco) and a continuous constraint solver (Ibex).

6.4. Learning Constraint Models

Participants: Nicolas Beldiceanu, Naina Razakarison.

- In the context of learning parametrized constraint models for highly structured problems we address in [38] the problem of finding the coefficients of polynomials in several variables from example parameter and function values.
- In the context of semi structured time series where the structural aspect is related to technological constraints we deal in [24] with the problem of extracting functional dependency constraints. The problem is motivated by extracting constraints from electricity production time series and is characterized by a larger set of samples (from 7 years and from 300 plants).

6.5. Meta Heuristics

Participants: Alejandro Reyes Amaro, Eric Monfroy, Florian Richoux, Charlotte Truchet.

- The aim is to develop and implement new algorithmical methods for constraint problems on massively parallel machines. We also conduct more theoretical studies about the parallelization of constraint problems. This year, we proposed a fairly sharp model to predict parallel speed-ups one can expect while parallelizing by a multi-walk parallel scheme any Las Vegas algorithm by just studying the distribution of sequential run-times [41]. This model shows a divergence of only 20% when predicting speed-ups over 256 cores, on very different benchmarks.
- To evaluate the scalability and parallelization of local search algorithms for SAT, [23] presents a statistical method based on the analysis of the runtime behavior of its sequential version.
- [26] and [26] deals with the use of metaheuristics for solving the resource constrained scheduling problem and the set covering problems.

6.6. Search and modelling

Participants: Eric Monfroy, Thierry Petit.

- In the context of autonomous search, [21] deals with the problem of automatically tuning a search strategy (i.e., variable value selection). For this purpose it uses so called *choice functions* which provide an evaluation of a strategy in term of a set of indicators. [36] and [31] go one step further by providing tuning and adaptation facilities at the level of the different components of a constraint solver.
- Using the MiniZinc modeling language, [32] shows how to model and solve the portfolio selection problem with constraint programming. Since more than ten year constraints for which the set of solutions can be matched to the language accepted by an automaton were introduced in many solvers (e.g., Choco, Gecode, SICStus). [40] describes an interface for describing such constraints in a more convenient way.
- Many discrete optimization problems have constraints on the objective function. Being able to represent such constraints is fundamental to deal with many real world industrial problems. In this work, we go one step further in the concept of topologically concentrate high values in a sequence of cost variables. We refine the work we previously published in CP2012 thanks to three generalizations of the focus constraint. We experiment successfully the technique in scheduling, round-robin and musical benchmarks. This work has been published at IJCAI 2013 [37].

6.7. Miscellaneous

Participants: Eric Monfroy, Florian Richoux.

- [15] gives a complete characterization of the complexity of the existential positive first-order logic, that one can interpret as model checking on monotone csp. We exhibit a dichotomy criterion remaining the same on finite domains of every cardinality, as well as countable and uncountable infinite domains.
- We develop an artificial intelligence, AIUR, to play the real time strategy game $StarCraft^{tm}$, using both machine learning and constraint-based techniques. AIUR finished 3$^{\text{rd}}$ to $StarCraft^{tm}$ AI competitions organized at the conferences AIIDE 2013 and CIG 2013. [18] presents a survey on AI techniques applied on $StarCraft^{tm}$.

TEXMEX Project-Team

6. New Results

6.1. Description of multimedia content

6.1.1. Multiscale image representations with component trees

Participants: Petra Bosilj, Ewa Kijak.

Joint work with Sébastien Lefevre, IRISA/SEASIDE, France.

The goal of this work is to study deeply the use of component trees, which aim at representing an image by the regions it contains at various scales through a tree-based structure, and their ability in the context of content-based image indexing and retrieval. Their invariance properties and their robustness to noise have motivated recent work in image indexing [83], [97], [98], but their usage in this field stays limited. The first part of this work was mainly dedicated to the study of various existing hierarchical representations. This leads to the presentation of a technique that arranges the elements of hierarchical representations of images according to a coarseness attribute [24]. The transformation is similar to filtering a hierarchy with a non-increasing attribute, and includes the results of multiple simple filterings with an increasing attribute. The transformed hierarchy can be then used for search space reduction prior to the image analysis process because it allows for direct access to the hierarchy elements at the same scale or a narrow range of scales.

6.1.2. Image representation

Participants: Rachid Benmokhtar, Jonathan Delhumeau, Guillaume Gravier, Philippe-Henri Gosselin, Hervé Jégou, Wanlei Zhao.

Partially in collaboration with Patrick Pérez, Technicolor, France.

Recent work on image retrieval have proposed to index images by compact representations encoding powerful local descriptors, such as the closely related vector of aggregated local descriptors (VLAD) and Fisher vector (FV). By combining them with a suitable coding technique, it is possible to encode an image in a few dozen bytes while achieving excellent retrieval results. We have pursed the research on this line of research by proposing two complementary contributions.

In [30], we revisited some assumptions proposed in this context regarding the handling of ”visual burstiness”, and shows that ad-hoc choices are implicitly done which are not desirable. Focusing on VLAD without loss of generality, we propose to modify several steps of the original design. Albeit simple, these modifications significantly improve VLAD and make it compare favorably against the state of the art.

In [65], we proposed a pooling strategy for local descriptors to produce a vector representation that is orientation-invariant yet implicitly incorporates the relative angles between features measured by their dominant orientation. This pooling is associated with a similarity metric that ensures that all the features have undergone a comparable rotation. This approach is especially effective when combined with dense oriented features, in contrast to existing methods that either rely on oriented features extracted on key points or on non-oriented dense features. The interest of our approach in a retrieval scenario is demonstrated on popular benchmarks comprising up to 1 million database images.

In [22], we propose to reduce the dimensionality of visual features for image categorization. We iteratively select sets of projections from an external dataset, using Bagging and feature selection thanks to SVM normals. Features are selected using weights of SVM normals in orthogonalized sets of projections. The bagging strategy is employed to improve the results and provide more stable selection. The overall algorithm linearly scales with the size of features, and is thus able to process large state-of-the-art image representations. Given Spatial Fisher Vectors as input, our method consistently improves the classification accuracy for smaller vector dimensionality, as demonstrated by our results on the popular and challenging PASCAL VOC 2007 benchmark.

6.1.3. Video classification

Participants: Kleber Jacques Ferreira de Souza, Guillaume Gravier, Philippe-Henri Gosselin.

In collaboration with Silvio Jamil F. Guimarães, PUC Minas, Brazil.

Most current motion descriptors for video classification are based on simple video segments, such as rectangular space-time blocks, or more recently rectangular space blocks that follow local trajectories. The aim of this study is to consider more complex video segments that better fit space-time elements of videos, thanks to recent methods for video segmentation proposed by S. Guimarães et al. These methods combine at the same time a fast extraction and stable regions, two essential properties for video indexing. The computation of local motion descriptors on these video segments lead to better video classification for human action recognition, when compared to current video indexing techniques.

6.1.4. Geo-localization of videos with multi-modality

Participants: Jonathan Delhumeau, Guillaume Gravier, Hervé Jégou.

Joint work with Michele Trevisiol, Yahoo! Labs, Spain, who visited the team in 2012.

Geotagging is the process of automatically adding geographical identification metadata to media objects, in particular to images and videos. In [63], we present a strategy to identify the geographic location of videos. First, it relies on a multi-modal cascade pipeline that exploits the available sources of information, namely the user upload history, his social network and a visual-based matching technique. Second, we present a novel divide & conquer strategy to better exploit the tags associated with the input video. It pre-selects one or several geographic area of interest of higher expected relevance and performs a deeper analysis inside the selected area(s) to return the coordinates most likely to be related to the input tags. The experiments were conducted as part of the MediaEval 2012 Placing Task, where we obtained the best results among the competitors when using no external information, i.e. not using any gazetteers nor any other kind of external information.

6.1.5. Violent keysound detection with audio words and Bayesian networks

Participants: Guillaume Gravier, Patrick Gros, Cédric Penet.

Joint work with Claire-Hélène Demarty, Technicolor, France.

We investigated a novel use of the well known audio words representations to detect specific audio events, namely gunshots and explosions, in order to get more robustness towards soundtrack variability in Hollywood movies [51]. An audio stream is processed as a sequence of stationary segments. Each segment is described by one or several audio words obtained by applying product quantization to standard features. Such a representation using multiple audio words constructed via product quantisation is one of the novelties described in this work. Based on this representation, Bayesian networks are used to exploit the contextual information in order to detect audio events. Experiments are performed on a comprehensive set of 15 movies, made publicly available. Results are comparable to the state of the art results obtained on the same dataset but show increased robustness to decision thresholds, however limiting the range of possible operating points in some conditions. Late fusion provides a solution to this issue.

6.2. Large scale indexing and classification

6.2.1. Parallelism and distribution for very large scale content-based image retrieval

Participants: Gylfi Gudmundsson, Diana Moise, Denis Shestakov, Laurent Amsaleg.

Two observations drove the design of the high-dimensional indexing technique developed in the framework of the Ph. D. thesis of Gylfli Gudmundson. Firstly, the collections are so huge, typically several terabytes, that they must be kept on secondary storage. Addressing disk related issues is thus central to our work. Secondly, all CPUs are now multi-core and clusters of machines are a commonplace. Parallelism and distribution are both key for fast indexing and high-throughput batch-oriented searching.

We developed a high-dimensional indexing technique called eCP. Its design includes the constraints associated to using disks, parallelism and distribution. At its core is an non-iterative unstructured vectorial quantization scheme. eCP builds on an existing indexing scheme that is main memory oriented. The first contribution in eCP is a set of extensions for processing very large data collections, reducing indexing costs and best using disks. The second contribution proposes multi-threaded algorithms for both building and searching, harnessing the power of multi-core processors. Datasets for evaluation contain about 25 million images or over 8 billion SIFT descriptors. The third contribution addresses distributed computing. We adapt eCP to the MapReduce programming model and use the Hadoop framework and HDFS for our experiments. This time we evaluate eCP's ability to scale-up with a collection of 100 million images, more than 30 billion SIFT descriptors, and its ability to scale-out by running experiments on more than 100 machines.

6.2.2. Contributions in image indexing

Participants: Hervé Jégou, Giorgos Tolias.

Partially in collaboration with Yannis Avrithis, National Technical University of Athens, Greece, Cai-Zhi Zhu and Shin'ichi Satoh, Japanese National Institute of Informatics, Japan.

In [62], we have considered a framework and its associated family of metrics to compare images based on their local descriptors. It encompasses the VLAD descriptor and matching techniques such as Hamming embedding. Making the bridge between these approaches leads us to propose a match kernel that takes the best of existing techniques by combining an aggregation procedure with a selective match kernel. Finally, the representation underpinning this kernel is approximated, providing a large scale image search both precise and scalable, as shown by our experiments on several benchmarks. We give a Matlab package associated with the paper that allows to reproduce the results of the most interesting variant.

On the same topic, we propose in [78] a query expansion technique for image search that is faster and more precise than the existing ones. An enriched representation of the query is obtained by exploiting the binary representation offered by the Hamming embedding image matching approach: The initial local descriptors are refined by aggregating those of the database, while new descriptors are produced from the images that are deemed relevant. This approach has two computational advantages over other query expansion techniques. First, the size of the enriched representation is comparable to that of the initial query. Second, the technique is effective even without using any geometry, in which case searching a database comprising 105k images typically takes 79 ms on a desktop machine. Overall, our technique significantly outperforms the visual query expansion state of the art on popular benchmarks. It is also the first query expansion technique shown effective on the UKB benchmark, which has few relevant images per query.

Finally, in [67] we considered a problem related to object retrieval, where we aim at retrieving, from a collection of images, all those in which a given query object appears. This problem is inherently asymmetric: the query object is mostly included in the database image, while the converse is not necessarily true. However, existing approaches mostly compare the images with symmetrical measures, without considering the different roles of query and database. This paper first measures the extent of asymmetry on large-scale public datasets reflecting this task. Considering the standard bag-of-words representation, we then propose new asymmetrical dissimilarities accounting for the different inlier ratios associated with query and database images. These asymmetrical measures depend on the query, yet they are compatible with an inverted file structure, without noticeably impacting search efficiency. Our experiments show the benefit of our approach, and show that the visual object retrieval task is better treated asymmetrically, in the spirit of state-of-the-art text retrieval.

6.2.3. Outlier detection applied to content-based image retrieval

Participants: Teddy Furon, Hervé Jégou.

The primary target of content based image retrieval is to return a list of images that are the most similar to a query image, which is usually done by ordering the images based on a similarity score. In most state-of-the-art systems, the magnitude of this score is very different from one query to another. This prevents us from making a proper decision about the correctness of the returned images. Our work [74] considers the applications where a confidence measurement is required, such as in copy detection or when a re-ranking stage is applied on a

short-list such as geometrical verification. For this purpose, we formulate image search as an outlier detection problem, and propose a framework derived from extreme values theory. We translate the raw similarity score returned by the system into a relevance score related to the probability that a raw score deviates from the estimated model of scores of random images. The method produces a relevance score which is normalized in the sense that it is more consistent across queries. Experiments performed on several popular image retrieval benchmarks and state-of-the-art image representations show the interest of our approach.

6.2.4. Exploiting motion characteristics for action classification in videos

Participants: Mihir Jain, Hervé Jégou.

In collaboration with Patrick Bouthemy, Inria/Serpico, France.

Several recent studies on action recognition have attested the importance of explicitly integrating motion characteristics in video description. In this work [43], we have re-visited the use of motion in videos, in order to better exploit it and improve action recognition systems. First, we established that adequately decomposing visual motion into dominant and residual motions, both in the extraction of the space-time trajectories and for the computation of descriptors, significantly improves action recognition algorithms. Then, we designed a new motion descriptor, the DCS descriptor, based on differential motion scalar quantities, divergence, curl and shear features. It captures additional information on the local motion patterns enhancing results. Finally, applying the recent VLAD coding technique proposed in image retrieval provides a substantial improvement for action recognition. Our three contributions are complementary and lead to significantly outperform all reported results on three challenging datasets, namely Hollywood 2, HMDB51 and Olympic Sports.

6.2.5. Recognizing events in videos

Participant: Hervé Jégou.

In collaboration with Matthijs Douze, Jérôme Revaud and Cordelia Schmid, Inria/LEAR, France.

We have addressed the problem of event retrieval for large-scale video collection. Given a video clip of a specific event, e.g., the wedding of Prince William and Kate Middleton, the goal is to retrieve other videos representing the same event from a dataset of over 100k videos.

Our first approach [55] encodes the frame descriptors of a video to jointly represent their appearance and temporal order. It exploits the properties of circulant matrices to compare the videos in the frequency domain. This offers a significant gain in complexity and accurately localizes the matching parts of videos. Furthermore, we extend product quantization to complex vectors in order to compress our descriptors, and to compare them in the compressed domain. Our method outperforms the state of the art both in search quality and query time on two large-scale video benchmarks for copy detection, Trecvid and CCweb. The evaluation has also been done on a new challenging dataset for event retrieval that we introduce: EVVE.

In a subsequent paper [39], we have made two other contributions to event retrieval in large collections of videos. First, we propose hyper-pooling strategies that encode the frame descriptors into a representation of the video sequence in a stable manner. Our best choices compare favorably with regular pooling techniques based on k-means quantization. Second, we introduce a technique to improve the ranking. It can be interpreted either as a query expansion method or as a similarity adaptation based on the local context of the query video descriptor. Experiments on public benchmarks show that our methods are complementary and improve event retrieval results, without sacrificing efficiency.

6.2.6. Large-scale SVM image classification

Participants: Thanh Nghi Doan, François Poulet.

Visual recognition remains an extremely challenging problem in computer vision research. Large datasets with millions images for thousands categories poses more challenges. We extend the state-of-the-art large scale linear classifier LIBLINEAR SVM and nonlinear classifier Power Mean SVM in two ways. The first one is to build a balanced bagging classifier with sampling strategy. The second one is to parallelize the training process of all binary classifiers with several multi-core computers [35]. We also applied the same approach to the stochastic gradient descent support vector machines (SVM-SGD) and to both state-of-the-art large linear classifier LIBLINEAR-CDBLOCK and nonlinear classifier Power Mean SVM in an incremental and parallel way [36].

6.2.7. Video copy detection with SNAP, a DNA indexing algorithm

Participants: Laurent Amsaleg, Guillaume Gravier.

In collaboration with Leonardo S. De Oliveira, Zenilton Kleber G. Do Patrocínio Jr. and Silvio Jamil F. Guimarães, PUC Minas, Brazil.

Near-duplicate video sequence identification consists in identifying real positions of a specific video clip in a video stream stored in a database. To address this problem, we proposed a new approach based on a scalable sequence aligner borrowed from proteomics [79]. Sequence alignment is performed on symbolic representations of features extracted from the input videos, based on an algorithm originally applied to bio-informatics. Experimental results demonstrate that our method performance achieved 94 % recall with 100 % precision, with an average searching time of about 1 second.

6.3. Security of multimedia contents and applications

6.3.1. Approximate nearest neighbors search with security and privacy requirements

Participants: Benjamin Mathon, Laurent Amsaleg, Teddy Furon.

In collaboration with Julien Bringer, Morpho, France.

This work presents a moderately secure but highly scalable and fast approximate nearest neighbors search. Our philosophy is to start from a state-of-the-art technique in this field based on approximate metrics: Euclidean distance based search in [47], [70], and cosine similarity based search in [42]. We then analyze the threats, and patch them avoiding as much as possible bricks penalizing too much the scalability and the speed. On the other hand, we do not completely prevent the players to infer some knowledge, but these limitations are well explained and experimentally assessed. The experimental body uses database of size much bigger than what the past secure solutions can handle.

6.3.2. A privacy-preserving framework for large-scale content-based information retrieval

Participants: Ewa Kijak, Laurent Amsaleg, Teddy Furon.

In close cooperation with Stéphane Marchand-Maillet, Li Weng and April Morton, University of Geneva, Switzerland.

We propose a privacy protection framework for large-scale content-based information retrieval. It offers two layers of protection. First, robust hash values are used as queries instead of original content or features. Second, the client can choose to omit certain bits in a hash value to further increase the ambiguity for the server. Due to the reduced information, it is computationally difficult for the server to know the client's interest. The server has to return the hash values of all possible candidates to the client. The client performs a search within the candidate list to find the best match. Since only hash values are exchanged between the client and the server, the privacy of both parties is protected.

We introduce the concept of *tunable privacy*, where the privacy protection level can be adjusted according to a policy. It is realized through hash-based piece-wise inverted indexing. The idea is to divide a feature vector into pieces and index each piece with a sub-hash value. Each sub-hash value is associated with an inverted index list.

The framework has been extensively tested using a large image database. We have evaluated both retrieval performance and privacy-preserving performance for a particular content identification application. Two different constructions of robust hash algorithms are used. One is based on random projections; the other is based on the discrete wavelet transform. Both algorithms exhibit satisfactory performance in comparison with state-of-the-art reference schemes. The results show that the privacy enhancement slightly improves the retrieval performance.

We consider the *majority voting attack* for estimating the query category and ID. Experiment results show that this attack is a threat when there are near-duplicates, but the success rate decreases with the number of omitted bits and the number of distinct items.

6.3.3. Privacy preserving data aggregation and service personalization using highly-scalable indexing techniques

Participants: Raghavendran Balu, Laurent Amsaleg, Hervé Jégou, Teddy Furon.

In collaboration with Armen Aghasaryan, Dimitre Davidov and Makram Bouzid, Alcatel-Lucent, and Sébastien Gambs, Inria/CIDRE, in the framework of the Alcaltel-Lucent / Inria common Lab.

A challenging approach to the problem of privacy preserving data aggregation and service personalization has recently been proposed in Bell Labs, which introduces a privacy-preserving intermediation layer between end-users and service providers. It uses a distributed variant of a Locality Sensitive Hashing (LSH) techniques of doing scalable nearest-neighbor search, adapted in a novel way, to discover similar users while preserving their privacy. This approach faces however several important challenges that will be targeted in the scope of this collaboration. The challenges are:

- *LSH optimization:* Definitions of hash functions as well as various LSH parameters need to be automatically tuned in order to achieve a good quality of generated recommendations with an expected level of the procured user anonymity. An interesting issue is the possibility of supervised machine learning. If some public profiles are available, more efficient clustering methods boost the quality of the recommendation service but their levels of anonymity have never been assessed so far.
- *Irreversibility of anonymization:* This needs to be evaluated for different attack models, e.g. exploiting the knowledge of LSH hashing functions or any other publically available information on users. It is equivalent as being able to define the region of the super high-dimensional space mapped into the same hashing results. This attack is bound to fail as this region is too large to leak information. However, the prior knowledge about the sparseness of the profiles might drastically reduce this region, and hence weaken the privacy.
- *System dynamics:* Dealing with the cold-start problem or controlling the dynamics of a running system when the profiles and the cluster assignments evolve over the time is yet another challenge this approach is confronted with. If these temporal issues are well studied in conventional relational databases, no clear solution is efficient in the recommendation area, and a fortiori in privacy enhancing recommendation systems.

6.4. Structuring multimedia content and summarization

6.4.1. Stream labeling for TV Structuring

Participants: Vincent Claveau, Guillaume Gravier, Patrick Gros, Emmanuelle Martienne, Abir Ncibi.

In this application, we focus on the problem of labeling the segments of TV streams according to their types (eg. programs, commercial breaks, sponsoring...). During this year, following the work initiated in 2012, we have proposed an in-depth analysis of the use of conditional random fields (CRF) for our task [50]. Through several experiments conducted on real TV streams, we have shown that the CRF yields high results compared with state-of-the-art approaches. In particular, CRF offers several ways to efficiently take the sequenciality of our stream labeling problem into account. We also showed that it is robust when dealing with few training data or few features.

6.4.2. Statistical tests for repetition detection in TV streams

Participant: Patrick Gros.

Detecting all repeated sequences in a TV stream is the first step of all techniques of TV stream structuring. We have improved our technique in several ways. First, a statistical hypothesis test with a corrected risk of Bonferroni was used to clean the repetitions of small sequences. Second, a content-based test is used to clean the remaining sequences, but also to complete the repeated sequences to their maximal length. One of our objective is to reduce the number of descriptor needed to achieve this task, given that this computation is the most expensive of the method. As a matter of fact, the method required computing the descriptors of 15.4 % of the images only.

6.4.3. Video summarization with constraint programming

Participants: Mohamed-Haykel Boukadida, Patrick Gros.

Joint work with Sid-Ahmed Berrani, Orange labs.

Up to now, most video summarization methods are based on concepts like saliency and often use a single modality. In order to develop a more general framework, we propose to use a constraint programming approach, where summarizing a video is seen as a constraint resolution problem, which consists in choosing certain excerpts with respect to various criteria. This year we studied several ways to model the problem in order to gain a maximum flexibility in the summary. A first model was based on the selection of shots, the second one on the selection of parts of shots; The third one does not relies on shots and select image sequences directly. The challenge is to express the useful constraints with these models and the limited possibilities of the solver.

6.4.4. Transcript-free spoken content summarization using motif discovery

Participants: Sébastien Campion, Guillaume Gravier.

Joint work with Frédéric Bimbot and Nathan Souviráa-Labastié, Inria/PANAMA, France.

Exploiting previous results on the unsupervised discovery of repeating words in speech signals, we proposed a method dedicated to transcript-free spoken content summarization. Extractive summarization is performed by selecting a small number of segments, typically one or two, which contains most of the repeated fragments [77]. Audio summaries were included in the Texmix demonstration and are currently being evaluated.

6.4.5. TV program structure discovery using grammatical inference

Participants: Guillaume Gravier, Bingqing Qu.

Joint work with Félicien Vallet and Jean Carrive, Institut National de l'Audiovisuel.

Video structuring, in particular applied to TV programs which have strong editing structures, mostly relies on supervised approaches either to retrieve a known structure for which a model has been obtained or to detect key elements from which a known structure is inferred. We investigated an unsupervised approach to recurrent TV program structuring, exploiting the repetitiveness of key structural elements across episodes of the same show. We cast the problem of structure discovery as a grammatical inference problem and show that a suited symbolic representation can be obtained by filtering generic events based on their reoccurring property [92]. The method follows three steps: *i*) generic event detection, *ii*) selection of events relevant to the structure and *iii*) grammatical inference from a symbolic representation. Experimental evaluation is performed on three types of shows, viz., game shows, news and magazines, demonstrating that grammatical inference can be used to discover the structure of recurrent programs with very limited supervision.

6.4.6. Discovering and linking related images in large collections

Participants: Guillaume Gravier, Hervé Jégou, Wanlei Zhao.

We have tackled the problem of image linking. One of the most successful method to link all similar images within a large collection is min-Hash, which is a way to significantly speed-up the comparison of images when the underlying image representation is bag-of-words. However, the quantization step of min-Hash introduces important information loss. In [66], we proposed a generalization of min-Hash, called Sim-min-Hash, to compare sets of real-valued vectors. We demonstrated the effectiveness of our approach when combined with the Hamming embedding similarity. Experiments on large-scale popular benchmarks demonstrated that Sim-min-Hash is more accurate and faster than min-Hash for similar image search. Linking a collection of one million images described by 2 billion local descriptors is done in 7 minutes on a single core machine.

6.5. Natural language processing in multimedia data

6.5.1. Text detection in videos

Participants: Khaoula Elagouni, Pascale Sébillot.

Texts embedded in multimedia documents often provide high level semantic clues that can be used in several applications or services. We thus aim at designing efficient Optical Character Recognition (OCR) systems able to recognize these texts. During the last three years, we have proposed three novel approaches, robust to text variability (different fonts, colors, sizes, etc.) and acquisition conditions (complex background, non-uniform lighting, low resolution, etc.). The first approach relies on a segmentation step and computes nonlinear separations between characters well adapted to the local morphology of images. The two other ones, called segmentation-free approaches, avoid the segmentation step by integrating a multi-scale scanning scheme: The first one relies on a graph model, while the second one uses a particular connectionist recurrent model able to handle spatial constraints between characters. In 2013, a precise evaluation and comparison between these approaches was conducted and published in [16].

6.5.2. Combining lexical cohesion and disruption for topic segmentation

Participants: Guillaume Gravier, Pascale Sébillot, Anca-Roxana Simon.

Topic segmentation classically relies on one of two criteria, either finding areas with coherent vocabulary use or detecting discontinuities. We proposed a segmentation criterion combining both lexical cohesion and disruption, enabling a trade-off between the two [58]. We provide the mathematical formulation of the criterion and an efficient graph based decoding algorithm for topic segmentation. Experimental results on standard textual data sets and on a more challenging corpus of automatically transcribed broadcast news shows demonstrate the benefit of such a combination. Gains were observed in all conditions, with segments of either regular or varying length and abrupt or smooth topic shifts. Long segments benefit more than short segments.However the algorithm has proven robust on automatic transcripts with short segments and limited vocabulary reoccurrences.

6.5.2.1. Information extraction and text mining

Participants: Vincent Claveau, Marie Béatrice Arnulphy.

Following the work initiated in the previous period, we have kept on working on relation extraction. During this year, we have proposed a new prototype that still relies on a supervised machine learning approach but we now rely on the sequence built from the shortest syntactic path between the entities, as it is done in many studies. These paths of lemmas are then used in a kNN whose similarity score is based on language modeling techniques. Based on this new prototype, we have participated to several tracks of the BioNLP challenges concerning the automatic extraction of relations in a specialized corpus. Results obtained with this simple and non-domain specific technique were relatively good, with a second and fourth ranks among the participants for the two tasks concerned [26].

We also pursued previous work on supervised techniques for entity extraction and classification. Instead of working on complex machine learning approaches, we rather use simple methods but the focus is set on clever similarity computing between training examples and candidates for which we make the most of existing information retrieval techniques. Our approach has been evaluated through our participation to BioNLP-ST13 competition, where it has been ranked first [26].

We have also proposed unsupervised techniques for knowledge discovery, more precisely, to bring out coherent groups of entities. Existing techniques are usually based on clustering; the challenge is then to define a notion of similarity between the relevant entities. In this work, we have proposed to divert conditional random fields (CRF) in order to calculate indirectly the similarities among text sequences. Our approach consists in generating artificial labeling problems on the data to be processed to reveal regularities in the labeling of the entities. The good results obtained shows the validity of our approach [27] and opens many research avenues for other knowledge discovery tasks.

6.5.3. *Unsupervised approaches to fine-grained morphological analysis*

Participants: Vincent Claveau, Ewa Kijak.

Following the work initiated in the previous years, we have kept on studying fine-grained morphological analysis for biomedical information retrieval. In the biomedical field, the key to access information is the use of specialized terms (like *photochemotherapy*). These complex morphological structures may prevent a user querying for *gastrodynia* to retrieve texts containing *stomachalgia*. The original unsupervised technique proposed in 2012 has been further developed and tested. In particular, during this year, we have shown that it largely outperforms state-of-the-art tools (*e.g.*, Morfessor and Derif) for morphological segmentation tasks. It also offers indirect morpho-lexical resources that are more reliable than hand-coded ones used in most state-of-the-art tools [11].

6.5.4. *Tree-structured named entities recognition*

Participants: Christian Raymond, Davy Weissenbacher.

Many natural language processing tasks needs the production of tree-structured outputs, like syntactic parsing, named entities recognition or language understanding. Currently, only machine learning based systems are robust enough to process the raw and noisy automatic transcribed speech while no machine learning paradigm are able to learn directly the tree structure in a reasonable time. In this work, we studied a solution to tackle the problem of predicting tree structured named entities from speech contents. We investigate a fast and robust decomposition strategy that was implemented and ranked best at the ETAPE NER evaluation campaign with results far better than those of the other participant systems [54].

6.5.5. *Fast machine learning algorithm for efficient combination of various features*

Participant: Christian Raymond.

Currently, in the field of natural language processing the machine learning algorithm "boosting over decision stumps" is often designed as the best off-the-shell classifier. It's actually widely used for his abilities to work on relatively big dataset, to operate intrinsically feature selection and to produce very good decision rules. We investigated a slight modification of this algorithm where the decision stumps are replaced by bonsai trees. Bonsai trees are small decision trees (with low depth) that can capture some structure in the data that decision stumps can not. This modification allows the boosting algorithm to exhibits better (or in the worst case similar) performances with a lower number of iteration the original algorithm needs. Thus allows in some cases a big improvement in term of performance for a lower cost in term of learning time. An application on image processing (typed/hand classification) exhibited interesting results in [94]

6.6. Competitions and international evaluation benchmarks

6.6.1. *FGcomp'2013, in conjunction with Imagenet*

Participants: Philippe-Henri Gosselin, Hervé Jégou.

Joint participation with Naila Murray and Florent Perronnin, Xerox Research Center Europe.

We have participated the the FGCOMP'2013 challenge and obtained the best results among all participants, see http://sites.google.com/site/fgcomp2013 Although the proposed system follows most of the standard Fisher classification pipeline, we have evaluated and used several key features and good practices that improve the accuracy when specifically considering fine-grained classification tasks [75]. In particular, we consider the late fusion of two systems both based on Fisher vectors, but that employ drastically different design choices that make them very complementary. Moreover, we show that a simple yet effective filtering strategy significantly boosts the performance for several class domains. The method is described in a technical report.

6.6.2. Hyperlink generation in broadcast videos

Participants: Guillaume Gravier, Pascale Sébillot, Anca-Roxana Simon.

Joint participation with Camille Guinaudeau, Heidelberg Institute of Technology (currently LIMSI-CNRS).

Following up on our 2012 participation, we participated in the Search and hyperlinking task implemented in the framework of the Mediaeval 2013 benchmark initiative. We limited ourselves to hyperlink generation, building on research results in natural language processing, information retrieval and topic segmentation, focusing our contribution on the selection of precise target segments for hyperlinks.

6.6.3. Maurdor campaign

Participant: Christian Raymond.

Joint participation with Yann Ricquebourg, Baptiste Poirriez, Aurélie Lemaitre and Bertrand Coüasnon, IRISA/Intuidoc.

We are participating to the ongoing MAURDOR campaign http://www.maurdor-campaign.org which aims at evaluating systems for automatic processing of written documents. The contribution of TEXMEX comes from the machine learning system based on boosting over bonsai trees we implemented. In the context of this campaign, we investigate the usefulness of this algorithm to combine efficiently features on a relatively big dataset. The very first result shows that this system get state-of-the-art performance while it is much faster than traditional SVM approaches.

6.6.4. Information extraction challenge at BioNLP-ST13

Participant: Vincent Claveau.

BioNLP Shared Task is a community-wide effort to address fine-grained, structural information extraction from biomedical literature. This year, several tasks were proposed and 22 teams participated. TexMex has proposed runs for three main tasks concerning entity extraction and categorization, and relation extraction. The methods proposed by our team are based on machine learning and information retrieval components. Although they do not exploit specialized or domain-specific knowledge, we obtained good results and ranked first, first and third according to the tasks.

TRISKELL Project-Team

6. New Results

6.1. Support for Reverse Engineering and Maintaining Feature Models

Feature Models (FMs) are a popular formalism for modelling and reasoning about commonality and variability of a system. In essence, FMs aim to define a set of valid combinations of features, also called configurations. In [35], we tackle the problem of synthesising an FM from a set of configurations. The main challenge is that numerous candidate FMs can be extracted from the same input configurations, yet only a few of them are meaningful and maintainable. We first characterise the different meanings of FMs and identify the key properties allowing to discriminate between them. We then develop a generic synthesis procedure capable of restituting the intended meanings of FMs based on inferred [72] or user-specified knowledge. Using tool support, we show how the integration of knowledge into FM synthesis can be realized in different practical application scenarios that involve reverse engineering and maintaining FMs.

6.2. Feature Model Extraction from Large Collections of Informal Product Descriptions

Feature Models (FMs) are used extensively in software product line engineering to help generate and validate individual product configurations and to provide support for domain analysis. As FM construction can be tedious and time-consuming, researchers have previously developed techniques for extracting FMs from sets of formally specified individual configurations, or from software requirements specifications for families of existing products. However, such artifacts are often not available. In [44] we present a novel, automated approach for constructing FMs from publicly available product descriptions found in online product repositories and marketing websites such as SoftPedia and CNET. While each individual product description provides only a partial view of features in the domain, a large set of descriptions can provide fairly comprehensive coverage. Our approach utilizes hundreds of partial product descriptions to construct an FM and is described and evaluated against antivirus product descriptions mined from SoftPedia.

6.3. On Product Comparison Matrices and Variability Models

Product comparison matrices (PCMs) provide a convenient way to document the discriminant features of a family of related products and now abound on the internet. Despite their apparent simplicity, the information present in existing PCMs can be very heterogeneous, partial, ambiguous, hard to exploit by users who desire to choose an appropriate product. Variability Models (VMs) can be employed to formulate in a more precise way the semantics of PCMs and enable automated reasoning such as assisted configuration. Yet, the gap between PCMs and VMs should be precisely understood and automated techniques should support the transition between the two. We propose variability patterns that describe PCMs content and conduct an empirical analysis of 300+ PCMs mined from Wikipedia [62], we also identify the limits of existing comparators, configurators and PCMs [67], [62]. Our findings are a first step toward better engineering techniques for maintaining and configuring PCMs.

6.4. Generating Counterexamples of Model-based Software Product Lines: An Exploratory Study

Model-based Software Product Line (MSPL) engineering aims at deriving customized models corresponding to individual products of a family. The design space of an MSPL is extremely complex to manage for the engineer, since the number of variants may be exponential and the derived product models have to conform to numerous well-formedness and business rules. We provide a way to generate MSPLs, called counterexamples, that can produce invalid product models despite a valid configuration in the variability model [49]. We provide a systematic and automated process, based on the Common Variability Language (CVL), to randomly search the space of MSPLs for a specific formalism. We validate the effectiveness of this process for three formalisms at different scales (up to 247 metaclasses and 684 rules).

6.5. Composing your Compositions of Variability Models

Modeling and managing variability is a key activity in a growing number of software engineering contexts. Support for composing variability models is arising in many engineering scenarios, for instance, when several subsystems or modeling artifacts, each coming with their own variability and possibly developed by different stakeholders, should be combined together. We consider in [34] the problem of composing feature models (FMs), a widely used formalism for representing and reasoning about a set of variability choices. We show that several composition operators can actually be defined, depending on both matching/merging strategies and semantic properties expected in the composed FM. We present four alternative forms and their implementations. We discuss their relative trade-offs w.r.t. reasoning, customizability, traceability, composability and quality of the resulting feature diagram. We summarize these findings in a reading grid which is validated by revisiting some relevant existing works. Our contribution should assist developers in choosing and implementing the right composition operators.

6.6. Extraction and Evolution of Architectural Variability Models in Plugin-based Systems

Variability management is a key issue when building and evolving software-intensive systems, making it possible to extend, configure, customize and adapt such systems to customers' needs and specific deployment contexts. A wide form of variability can be found in extensible software systems, typically built on top of plugin-based architectures that offer a (large) number of configuration options through plugins. In an ideal world, a software architect should be able to generate a system variant on-demand, corresponding to a particular assembly of plugins. To this end, the variation points and constraints between architectural elements should be properly modeled and maintained over time (i.e., for each version of an architecture). A crucial, yet error-prone and time-consuming, task for a software architect is to build an accurate representation of the variability of an architecture, in order to prevent unsafe architectural variants and reach the highest possible level of flexibility. In [23], we propose a reverse engineering process for producing a variability model (i.e., a feature model) of a plugin-based architecture. We develop automated techniques to extract and combine different variability descriptions, including a hierarchical software architecture model, a plugin dependency model and the software architect knowledge. By computing and reasoning about differences between versions of architectural feature models, software architect can control both the variability extraction and evolution processes. The proposed approach has been applied to a representative, large-scale plugin-based system (FraSCAti), considering different versions of its architecture. We report on our experience in this context.

6.7. FAMILIAR: A Domain-Specific Language for Large Scale Management of Feature Models

The feature model formalism has become the de facto standard for managing variability in software product lines (SPLs). In practice, developing an SPL can involve modeling a large number of features representing di erent viewpoints, sub-systems or concerns of the software system. This activity is generally tedious and error-prone. In [24], we present FAMILIAR a Domain-Specific Language (DSL) that is dedicated to the large scale management of feature models and that complements existing tool support. The language provides a powerful support for separating concerns in feature modeling, through the provision of composition and decomposition operators, reasoning facilities and scripting capabilities with modularization mechanisms. We illustrate how an SPL consisting of medical imaging services can be practically managed using reusable FAMILIAR scripts that implement reasoning mechanisms. We also report on various usages and applications of FAMILIAR and its operators, to demonstrate their applicability to di erent domains and use for di erent purposes.

6.8. Web Configurators

Nowadays, mass customization has been embraced by a large portion of the industry. As a result, the web abounds with sales configurators that help customers tailor all kinds of goods and services to their specific

needs. In many cases, configurators have become the single entry point for placing customer orders. As such, they are strategic components of companies' information systems and must meet stringent reliability, usability and evolvability requirements. However, the state of the art lacks guidelines and tools for efficiently engineering web sales configurators. To tackle this problem, empirical data on current practice is required. The paper [51] reports on a systematic study of 111 web sales configurators along three essential dimensions: rendering of configuration options, constraint handling, and configuration process support. Based on this, we highlight good and bad practices in engineering web sales configurator. The reported quantitative and qualitative results open avenues for the elaboration of methodologies to (re-)engineer web sales configurators. In [48] we focus on how to associate product configurations to visual representations in a Web configurator. We present a formal statement of the problem and a model-driven perspective.

6.9. Separating Concerns in Feature Models

Feature models (FMs) are a popular formalism to describe the commonality and variability of a set of assets in a software product line (SPL). SPLs usually involve large and complex FMs that describe thousands of features whose legal combinations are governed by many and often complex rules. The size and complexity of these models is partly explained by the large number of concerns considered by SPL practitioners when managing and configuring FMs. In the chapter [68], we first survey concerns and their separation in FMs, highlighting the need for more modular and scalable techniques. We then revisit the concept of view as a simplified representation of an FM. We finally describe a set of techniques to specify, visualize and verify the coverage of a set of views. These techniques are implemented in complementary tools providing practical support for feature-based configuration and large scale management of FMs.

6.10. Bridging the Chasm between Executable Metamodeling and Models of Computation

The complete and executable definition of a Domain Specific Language (DSL) relies on the specification of two essential facets: a model of the domain-specific concepts with actions and their semantics; a scheduling model that orchestrates the actions of a domain-specific model. Metamodels can capture the former facet, while Models of Computation (MoCs) capture the latter facet. Unfortunately, theories and tools for metamodeling and MoCs have evolved independently, creating a cultural and technical chasm between the two communities. We introduce a new framework to bridge a metamodel and a MoC in a modular fashion [43]. This bridge allows (i) the complete and executable definition of a DSL, (ii) the reuse of MoCs for different domain-specific metamodels, and (iii) the use of different MoCs for a given metamodel, to cope with variation points of a DSL.

6.11. Reifying Concurrency for Executable Metamodeling

Current metamodeling techniques can be used to specify the syntax and semantics of domain specific modeling languages (DSMLs). However, there is currently very little support for explicitly specifying concurrency semantics using metamodels. We reify concurrency as a metamodeling facility, leveraging formalization work from the concurrency theory and models of computation (MoC) community [42]. The essential contribution of this paper is a proposed language workbench for binding domain-specific concepts and models of computation through an explicit event structure at the metamodel level. We illustrate these novel metamodeling facilities for designing two variants of a concurrent and timed final state machine, and provide other experiments to validate the scope of our approach.

6.12. Using Model Types to Support Contract-Aware Model Substitutability

Model typing brings the benefit associated with well-defined type systems to model-driven development (MDD) through the assignment of specific types to models. In particular, model type systems enable reuse of model manipulation operations (e.g., model transformations), where manipulations defined for models of a supertype can be used to manipulate models of subtypes. Existing model typing approaches are limited to

structural typing defined in terms of object-oriented metamodels (e.g., MOF) in which the only structural (well-formedness) constraints are those that can be expressed directly in metamodeling notations (e.g., multiplicity and element containment constraints). We propose an extension to model typing that takes into consideration structural invariants, other than those that can be expressed directly in metamodeling notation, and specifications of behaviors associated with model types [64]. The approach supports contract-aware substitutability, where contracts are defined in terms of invariants and pre-/postconditions expressed using OCL. Support for behavioral typing paves the way for behavioral substitutability. We also describe a technique to rigorously reason about model type substitutability as supported by contracts and apply the technique in use cases from the optimizing compiler community.

6.13. Variability Support in Domain-Specific Language Development

Domain Specific Languages (DSLs) are widely adopted to capitalize on business domain experiences. Consequently, DSL development is becoming a recurring activity. Unfortunately, even though it has its benefits, language development is a complex and time-consuming task. Languages are commonly realized from scratch, even when they share some concepts and even though they could share bits of tool support. This cost can be reduced by employing modern modular programming techniques that foster code reuse. However, selecting and composing these modules is often only within the reach of a skilled DSL developer. We propose to combine modular language development and variability management, with the objective of capitalizing on existing assets [63]. This approach explicitly models the dependencies between language components, thereby allowing a domain expert to configure a desired DSL, and automatically derive its implementation. The approach is tool supported, using Neverlang to implement language components, and the Common Variability Language (CVL) for managing the variability and automating the configuration. We illustrate our approach with the help of different case studies, including the implementation of a family of DSLs to describe variants of state machines.

6.14. Automatically Searching for Metamodel Well-Formedness Rules in Examples and Counter-Examples

Current metamodeling formalisms support the definition of a metamodel with two views: classes and relations, that form the core of the metamodel, and well-formedness rules, that constraints the set of valid models. While a safe application of automatic operations on models requires a precise definition of the domain using the two views, most metamodels currently present in repositories have only the first one part. We propose in [47] to start from valid and invalid model examples in order to automatically retrieve well-formedness rules in OCL using Genetic Programming. The approach is evaluated on metamodels for state machines and features diagrams. The experiments aim at demonstrating the feasibility of the approach and at illustrating some important design decisions that must be considered when using this technique.

6.15. Building Modular and Efficient DSLs: Mashup of Meta-Languages and its Implementation in the Kermeta Language Workbench

With the growing use of domain-specific languages (DSL) in industry, DSL design and implementation goes far beyond an activity for a few experts only and becomes a challenging task for thousands of software engineers. DSL implementation indeed requires engineers to care for various concerns, from abstract syntax, static semantics, behavioral semantics, to extra-functional issues such as run-time performance. We propose an approach that uses one meta-language per language implementation concern [27] in the new version (v2) of the Kermeta workbench. We show that the usage and combination of those meta-languages is simple and intuitive enough to deserve the term "mashup". We evaluate the approach by completely implementing the non trivial fUML modeling language, a semantically sound and executable subset of the Unified Modeling Language (UML) ; Kompren, a DSL for designing and implementing model slicers ; and KCVL, the Commun Variability Language dedicated to variability management in software design models [65].

6.16. On the Globalization of Modeling Languages

In the software and systems modeling community, research on domain-specific modeling languages (DSMLs) is focused on providing technologies for developing languages and tools that allow domain experts to develop system solutions efficiently in a particular domain. Unfortunately, the current lack of support for explicitly relating concepts expressed in different DSMLs makes it very difficult for software and system engineers to reason about information spread across models describing different system aspects. Supporting coordinated use of DSMLs leads to what we call the globalization of modeling languages. We present a research initiative that broadens the current DSML research focus beyond the development of independent DSMLs to one that provides support for globalized DSMLs, that is, DSMLs that facilitate coordination of work across different domains of expertise [31]. We explore this new grand challenge in recent workshops, *e.g.,* GlobalDSL'13 at ECSA, ECMFA and ECOOP 2013 [69], and GEMOC'13 at MODELS 2013 [70].

6.17. Automating the Maintenance of Non-functional System Properties using Demonstration-based Model Transformation

Given a base model with functional components, maintaining the non-functional properties that crosscut the base model has become an essential modeling task when using DSMLs. We present a demonstration-based approach to automate the maintenance of non-functional properties in DSMLs [29]. Instead of writing model transformation rules explicitly, users demonstrate how to apply the non-functional properties by directly editing the concrete model instances and simulating a single case of the maintenance process. An inference engine generates generic model transformation patterns, which can be refined by users and then reused to automate the same evolution and maintenance task in other models. Our demonstration-based approach has been applied to several scenarios, such as auto-scaling and model layout.

6.18. Improving Reusability and Automation in Software Process Lines

Software processes orchestrate manual or automatic tasks to create new software products that meet the requirements of specific projects. While most of the tasks are about inventiveness, modern developments also require recurrent, boring and time-consuming tasks (e.g., the IDE configuration, or the continuous integration setup). Such tasks struggle to be automated due to their various execution contexts according to the requirements of specific projects. We propose a methodology that benefits from an explicit modeling of a family of processes to identify the possible reuse of automated tasks in software processes [60]. Then, we propose a tool-supported approach that integrates both reuse and automation [61]. It consists of reusing processes from an SPL according to projects' requirements. The processes are bound to components that automate their execution. When the variability of a process to execute is not fully resolved, our approach consists of resolving this variability during the execution of this process. We illustrate our approach on industrial projects in a software company, as well as on a family of processes for designing and implementing modeling languages. Our approach promoted the identification of possible automated tasks for configuring IDEs and continuous integration, their reuse in various projects of the company, and the automation of their execution, while enabling to resolve process variability during the execution.

6.19. Towards Trust-Aware and Self-Adaptive Systems

The dynamic conditions under which Future Internet (FI) applications must execute call for self-adaptive software to cope with unforeseeable changes in the application environment. Software engineering currently provides frameworks to develop reasoning engines that support the runtime adaptation of distributed, heterogeneous applications. However, these frameworks have very limited support to address security concerns of these application, hindering their usage for FI scenarios. We address this challenge by enhancing self-adaptive systems with the concepts of trust and reputation [58]. Trust improves decision-making processes under risk and uncertainty, in turn improving security of self-adaptive FI applications.

6.20. SOA Antipatterns: an Approach for their Specification and Detection

The changes resulting from the evolution of Service Based Systems (SBSs) may degrade their design and quality of service (QoS) and may often cause the appearance of common poor solutions in their architecture, called antipatterns. We introduce a novel and innovative approach supported by a framework for specifying and detecting antipatterns in SBSs [25]. We specify ten well-known and common antipatterns, including Multi Service and Tiny Service, and automatically generate their detection algorithms. We validate the detection algorithms in terms of precision and recall on two systems developed independently. This validation demonstrates that our approach enables the specification and detection of SOA antipatterns with an average precision of 90% and recall of 97.5%.

6.21. Automated Measurement of Models of Requirements

One way to formalize system requirements is to express them using the object-oriented paradigm. In this case, the class model representing the structure of requirements is called a requirements metamodel, and requirements themselves are object-based models of natural-language requirements. We show that such object-oriented requirements are well-suited to support a large class of requirements metrics[28]. We define a requirements metamodel and use an automated measurement approach proposed in our previous work to specify requirements metrics. We show that it is possible to integrate 78 metrics from 11 different papers in the proposed framework. The software that computes the requirements metric values is fully generated from the specification of metrics.

6.22. Empirical Evidence of Large-Scale Diversity in API Usage of Object-Oriented Software

In this paper, we study how object-oriented classes are used across thousands of software packages. We concentrate on "usage diversity'", defined as the different statically observable combinations of methods called on the same object. We present empirical evidence that there is a significant usage diversity for many classes. For instance, we observe in our dataset that Java's String is used in 2460 manners. We discuss the reasons of this observed diversity and the consequences on software engineering knowledge and research [56].

6.23. Efficient high-level abstractions for web programming

Writing large Web applications is known to be difficult. One challenge comes from the fact that the application's logic is scattered into heterogeneous clients and servers, making it difficult to share code between both sides or to move code from one side to the other. Another challenge is performance: while Web applications rely on ever more code on the client-side, they may run on smart phones with limited hardware capabilities. These two challenges raise the following problem: how to benefit from high-level languages and libraries making code complexity easier to manage and abstracting over the clients and servers differences without trading this ease of engineering for performance? In [59], we present high-level abstractions defined as deep embedded DSLs in Scala that can generate efficient code leveraging the characteristics of both client and server environments. We compare performance on client-side against other candidate technologies and against hand written low-level JavaScript code. Though code written with our DSL has a high level of abstraction, our benchmark on a real world application reports that it runs as fast as hand tuned low-level JavaScript code.

6.24. Exploring Optimal Service Compositions in Highly Heterogeneous and Dynamic Service-Based Systems

Service-oriented pervasive systems, composed of a large number of devices with heterogeneous capabilities where devices' resources are abstracted as software services, challenge the creation of high-quality composite applications. Resource heterogeneity, dynamic network connectivity, and a large number of highly distributed service providers complicate the process of creating applications with specific QoS requirements. Existing approaches to service composition control the QoS of an application solely by changing the set of participating

concrete services which is not suitable for ad-hoc service-based systems characterised by high intermittent connectivity and resource heterogeneity. In [46], we propose a flexible way of formulating composition configurations suitable for such service-based systems. Our formulation proposes the combined consideration of the following factors that affect the QoS of a composed service: (a) service selection, (b) orchestration partitioning, and (c) orchestrator node selection. We show that the proposed formulation enables the definition of service composition configurations with 49% lower response time, 28% lower network latency, 36% lower energy consumption, and 13% higher success ratio compared to those defined with the traditional approach. In [45], we present the problem of efficiently exploring at runtime the serach space of possible configurations for a service orchestration with various Quality of Services.

VISAGES Project-Team

6. New Results

6.1. Image Computing: Detection, Segmentation, Registration and Analysis

6.1.1. A Mathematical Framework for the Registration and Analysis of Multi-Fascicle Models for Population Studies of the Brain Microstructure

Participant: Olivier Commowick.

Diffusion tensor imaging (DTI) is unable to represent the diffusion signal arising from multiple crossing fascicles and freely diffusing water molecules. Generative models of the diffusion signal, such as multi-fascicle models, overcome this limitation by providing a parametric representation for the signal contribution of each population of water molecules. These models are of great interest in population studies to characterize and compare the brain microstructural properties. Central to population studies is the construction of an atlas and the registration of all subjects to it. However, the appropriate definition of registration and atlasing methods for multi-fascicle models have proven challenging. This paper proposes [32] a mathematical framework to register and analyze multi-fascicle models. Specifically, we define novel operators to achieve interpolation, smoothing and averaging of multi-fascicle models. We also define a novel similarity metric to spatially align multi-fascicle models. Our framework enables simultaneous comparisons of different microstructural properties that are confounded in conventional DTI. The framework is validated on multi-fascicle models from 24 healthy subjects and 38 patients with tuberous sclerosis complex, 10 of whom have autism. We demonstrate the use of the multi-fascicle models registration and analysis framework in a population study of autism spectrum disorder.

6.1.2. Multimodal rigid-body registration of 3D brain images using bilateral symmetry

Participants: Sylvain Prima, Olivier Commowick.

In this paper we show how to use the approximate bilateral symmetry of the brain with respect to its interhemispheric fissure for intra-subject (rigid-body) mono- and multimodal 3D image registration. We propose to define and compute an approximate symmetry plane in the two images to register and to use these two planes as constraints in the registration problem. This 6-parameter problem is thus turned into three successive 3-parameter problems. Our hope is that the lower dimension of the parameter space makes these three subproblems easier and faster to solve than the initial one. We implement two algorithms to solve these three subproblems in the exact same way, within a common intensity-based framework using mutual information as the similarity measure. We compare this symmetry-based strategy with the standard approach (i.e. direct estimation of a 6-parameter rigid-body transformation), also implemented within the same framework, using synthetic and real datasets. We show in [44] our symmetry-based method to achieve subvoxel accuracy with better robustness and larger capture range than the standard approach, while being slightly less accurate and slower. Our method also succeeds in registering clinical MR and PET images with a much better accuracy than the standard approach. Finally, we propose a third strategy to decrease the run time of the symmetry-based approach and we give some ideas, to be tested in future works, on how to improve its accuracy.

6.1.3. Distortion Correction in EPI Diffusion Weighted Images

Participants: Renaud Hedouin, Olivier Commowick.

We have compared and developed several methods which correct distortion of EPI images. The most popular method field map do not give optimal results. We have implemented and improved a method based on reversed phase encoding gradient which give good results. To correct diffusion weighted images this method only need one reversed phase encoding gradient $B0$ image which not need substantial additional acquisition time.

6.1.4. Using bilateral symmetry to improve non-local means denoising of MR brain images

Participants: Sylvain Prima, Olivier Commowick.

The popular NL-means denoising algorithm proposes to modify the intensity of each voxel of an image by a weighted sum of the intensities of similar voxels. The success of the NL-means rests on the fact that there are typically enough such similar voxels in natural, and even medical images; in other words, that there is some self-similarity/redundancy in such images. However, similarity between voxels (or rather, between patches around them) is usually only assessed in a spatial neighbourhood of the voxel under study. As the human brain exhibits approximate bilateral symmetry, one could wonder whether a voxel in a brain image could be more accurately denoised using information from both ipsi- and contralateral hemispheres. This is the idea we have investigated in this paper [45]. We define and compute a mid-sagittal plane which best superposes the brain with itself when mirrored about the plane. Then we use this plane to double the size of the neighbourhoods and hopefully find additional interesting voxels to be included in the weighted sum. We evaluate this strategy using an extensive set of experiments on both simulated and real datasets.

6.1.5. Detection of Multiple Sclerosis Lesions using Dictionary Learning

Participants: Hrishikesh Deshpande, Pierre Maurel, Christian Barillot.

Multiple sclerosis (MS) is a chronic, autoimmune, inflammatory disease of the central nervous system, in which certain areas of brain develop MS lesions, which are characterized by demyelination. Over the last years, various models combined with supervised and unsupervised classification methods have been proposed for detection of MS lesions using magnetic resonance images. Recently, signal modeling using sparse representations (SR) has gained tremendous attention and is an area of active research. SR allows coding data as sparse linear combinations of the elements of over-complete dictionary and has led to interesting image recognition results. The dictionary used for sparse coding plays a key role in the classification process. In this work, we have proposed to learn class specific dictionaries and develop new classification scheme, to automatically detect MS lesions in 3-D multi-channel magnetic resonance images.

6.1.6. Multiple Sclerosis Lesion Detection in Clinically Isolated Syndromes

Participants: Yogesh Karpate, Olivier Commowick, Christian Barillot.

Quantitative assessment of Multiple Sclerosis Lesions (MSL) in Clinically Isolated Syndromes (CIS) is important, as they are a precursor to subsequent stages of the disease. We address the problem of lesion patch detection with respect to Normally Appearing Brain Tissues (NABT). Our approach consists in learning rotationally invariant MSL and NABT multimodal intensity signatures based on 3D spherical gabor descriptors. This learning step, done once and for all, is followed by a testing step for the patient patches with an exemplar SVM. First, we develop a framework for selecting focused region of interest (fROI) using linear SVM for scoring. This allows an excellent trade-off between speed and accuracy. Second, building rotational invariant and scale independent features for accurate representation of image signatures. The extracted features are sensitive to the orientation of the analyzed image. This is a drawback in classification and retrieval applications. We handle this problem by using shperical Gabor descriptors. And last, we apply max pooling for down sampling of feature vectors. For the classification purpose we use a standard linear Support Vector Machine(SVM). The main contribution of the work is to build binary classifier to discriminate NABTs and MSLs based upon robust image representation. We have validated our approach on synthetic and real patient data. The synthetic lesion data is generated with noise, without noise and with bias field. Further, validation is carried out in three different scenarios. First, we evaluate our classifier using K-fold started with cross validation using NABT from healthy volunteers and MSL from CIS patients, then the detection of NABT and MSL from CIS patients on known patches is performed. The last evaluation concerned the full search algorithm.

6.1.7. Intensity Normalization in Longitudinal MS Patients

Participants: Yogesh Karpate, Olivier Commowick, Christian Barillot.

This work proposes a longitudinal intensity normalization algorithm for multi-channel MRI of brain of MS patient in the presence of lesions, aiming towards stable and consistent longitudinal segmentation. This approach is parametric and developed using two different forms of Robust Expectation Maximization (EM). The first is Spatio-Temporal Robust Expected Maximization (STREM) and other being EM with beta divergence. We validated our method on real longitudinal multiple sclerosis subjects.

6.2. Image processing on Diffusion Weighted Magnetic Resonance Imaging

6.2.1. Statistical Analysis of White Matter Integrity for the Clinical Study of Specific Language Impairment in Children

Participants: Olivier Commowick, Camille Maumet, Aymeric Stamm, Jean-Christophe Ferré, Christian Barillot.

Children affected by Specific Language Impairment (SLI) fail to develop a normal language capability. To date, the etiology of SLI remains largely unknown. It induces difficulties with oral language which cannot be directly attributed to intellectual deficit or other developmental delay. Whereas previous studies on SLI focused on the psychological and genetic aspects of the pathology, few imaging studies investigated defaults in neuroanatomy or brain function. We have proposed [53] to investigate the integrity of white matter in SLI thanks to diffusion Magnetic Res- onance Imaging. An exploratory analysis was performed without a priori on the impaired regions. A region of interest statistical analysis was performed based, first, on regions defined from Catani's atlas and, then, on tractography-based regions. Both the mean fractional anisotropy and mean apparent diffusion coefficient were compared across groups. To the best of our knowledge, this is the first study focusing on white matter integrity in specific language impairment. 22 children with SLI and 19 typically developing children were involved in this study. Overall, the tractography-based approach to group comparison was more sensitive than the classical ROI-based approach. Group differences between controls and SLI patients included decreases in FA in both the perisylvian and ventral pathways of language, comforting findings from previous functional studies.

6.2.2. Adaptive Multi-modal Particle Filtering for Probabilistic White Matter Tractography

Participants: Aymeric Stamm, Olivier Commowick, Christian Barillot.

Particle filtering has recently been introduced to perform probabilistic tractography in conjunction with DTI and Q-Ball models to estimate the diffusion information. Particle filters are particularly well adapted to the tractography problem as they offer a way to approximate a probability distribution over all paths originated from a specified voxel, given the diffusion information. In practice however, they often fail at consistently capturing the multi-modality of the target distribution. For brain white matter tractography, this means that multiple fiber pathways are unlikely to be tracked over extended volumes. We have proposed [51] to remedy this issue by formulating the filtering distribution as an adaptive M-component non-parametric mixture model. Such a formulation preserves all the properties of a classical particle filter while improving multi-modality capture. We apply this multi-modal particle filter to both DTI and Q-Ball models and propose to estimate dynamically the number of modes of the filtering distribution. We show on synthetic and real data how this algorithm outperforms the previous versions proposed in the literature.

6.2.3. Tracking the Cortico-Spinal Tract from Low Spatial and Angular Resolution Diffusion MRI

Participants: Aymeric Stamm, Olivier Commowick, Christian Barillot.

We have participated to the annual MICCAI workshop on DTI tractography [52]. We presented a pipeline to reconstruct the corticospinal tract (CST) that connects the spinal cord to the motor cortex. The proposed method combines a new geometry-based multi-compartment diffusion model coined Diffusion Directions Imaging and a new adaptive multi-modal particle filter for tractography. The DTI Tractography challenge proposes to test our methods in the context of neurosurgical planning of tumor removal, where very low spatial and angular resolution diffusion data is available due to severe acquisition time constraints. We took

up the challenge and present our reconstructed CSTs derived from a single-shell acquisition scheme at b = 1000 s/mm2 with only 20 or 30 diffusion gradients (low angular resolution) and with images of 5 mm slice thickness (low spatial resolution).

6.3. Medical Image Computing in Brain Pathologies

6.3.1. Semi-Automatic Classification of Lesion Patterns in Patients with Clinically Isolated Syndrome

Participants: Olivier Commowick, Jean-Christophe Ferré, Gilles Edan, Christian Barillot.

Multiple sclerosis (MS) is neuro-degenerative disease of the Central Nervous System characterized by the loss of myelin. A Clinically Isolated Syndrome (CIS) is a first neurological episode caused by inflammation/demyelination in the central nervous system which may lead to MS. Better understanding of the disease at its onset will lead to a better discovery of pathogenic mechanisms, allowing suitable therapies at an early stage. We have proposed [37] an automatic segmentation algorithm for two different contrast agents, used within a framework for early characterization of CIS patients according to lesion patterns, and more specifically according to the nature of the inflammatory patterns of these lesions. We expect that the proposed framework can infer new prospective figures from the earliest imaging signs of MS since it can provide a classification of different types of lesions across patients. The lesion detection algorithm based on intensity normalization and subtraction of the used MRI data is a pivotal step, since it avoids the time-demanding task of manual delineation.

6.3.2. Multiple Sclerosis Lesions Evolution in Patients with Clinically Isolated Syndrome

Participants: Olivier Commowick, Jean-Christophe Ferré, Gilles Edan, Christian Barillot.

Multiple sclerosis (MS) is a disease with heterogeneous evolution among the patients. Some classifications have been carried out according to either the clinical course or the immunopathological profiles. Epidemiological data and imaging are showing that MS is a two-phase neurodegenerative inflammatory disease. At the early stage it is dominated by focal inflammation of the white matter (WM), and at a latter stage it is dominated by diffuse lesions of the grey matter and spinal cord. A Clinically Isolated Syndrome (CIS) is a first neurological episode caused by inflammation/demyelination in the central nervous system which may lead to MS. Few studies have been carried out so far about this initial stage. Better understanding of the disease at its onset will lead to a better discovery of pathogenic mechanisms, allowing suitable therapies at an early stage. We have proposed [36] a new data processing framework able to provide an early characterization of CIS patients according to lesion patterns, and more specifically according to the nature of the inflammatory patterns of these lesions. The method is based on a two layers classification. Initially, the spatio-temporal lesion patterns are classified using a tensor-like representation. The discovered lesion patterns are then used to identify group of patients and their correlation to 15 months follow-up total lesion loads (TLL), which is so far the only image-based figure that can potentially infer future evolution of the pathology. We expect that the proposed framework can infer new prospective figures from the earliest imaging sign of MS since it can provide a classification of different types of lesion across patients.

6.3.3. Arterial Spin Labeling at 3T in semantic dementia: perfusion abnormalities detection and comparison with FDG-PET

Participants: Isabelle Corouge, Jean-Christophe Ferré, Elise Bannier, Aymeric Stamm, Christian Barillot, Jean-Yves Gauvrit.

Arterial Spin Labeling (ASL) is a non invasive perfusion imaging technique which has shown great diagnosis potential in dementia. However, it has never been applied to semantic dementia (SD), a rare subtype of frontotemporal lobar degeneration characterized by the gradual loss of conceptual knowledge, which is actually explored by a now well established marker of SD: ^{18}fluorodeoxyglucose-positron emission tomography (FDG-PET) imaging. Although ASL and FDG-PET respectively measure perfusion and metabolism, they have

been shown to be strongly correlated. In this work, we explore the ability of ASL to detect perfusion abnormalities in SD in comparison with FDG-PET. Using patients and healthy subjects data from an ongoing clinical study, we apply our analysis framework starting with visual comparison of ASL and FDG-PET, and focusing on ASL data preprocessing and statistical analysis at the individual and group level. Preliminary results yield concordant observations between ASL and FDG-PET as well as expected hypoperfusions in SD, namely in the left temporal lobe, thus suggesting the potential of ASL to assess perfusion impairments in SD.

6.4. Vascular Imaging and Arterial Spin Labelling

6.4.1. Patient-specific detection of perfusion abnormalities combining within-subject and between-subject variances in Arterial Spin Labeling.

Participants: Camille Maumet, Pierre Maurel, Jean-Christophe Ferré, Christian Barillot.

In this paper, patient-specific perfusion abnormalities in Arterial Spin Labeling (ASL) were identified by comparing a single patient to a group of healthy controls using a mixed-effect hierarchical General Linear Model (GLM). Two approaches are currently in use to solve hierarchical GLMs: (1) the homoscedastic approach assumes homogeneous variances across subjects and (2) the heteroscedastic approach is theoretically more efficient in the presence of heterogeneous variances but algorithmically more demanding. In practice, in functional magnetic resonance imaging studies, the superiority of the heteroscedastic approach is still under debate. Due to the low signal-to-noise ratio of ASL sequences, within-subject variances have a significant impact on the estimated perfusion maps and the heteroscedastic model might be better suited in this context. In this paper we studied how the homoscedastic and heteroscedastic approaches behave in terms of specificity and sensitivity in the detection of patient-specific ASL perfusion abnormalities. Validation was undertaken on a dataset of 25 patients diagnosed with brain tumors and 36 healthy volunteers. We showed evidence of heterogeneous within-subject variances in ASL and pointed out an increased false positive rate of the homoscedastic model. In the detection of patient-specific brain perfusion abnormalities with ASL, modeling heterogeneous variances increases the sensitivity at the same specificity level [24].

6.4.2. An a contrario approach for the detection of activated brain areas in fMRI

Participants: Camille Maumet, Pierre Maurel, Jean-Christophe Ferré, Christian Barillot.

BOLD functional MRI (fMRI) is now a widespread imaging technique to study task-related activity in the brain. However, getting the areas of activation at the individual subject level is still an open issue. The standard massively univariate statistical analysis is usually performed after smoothing the data and makes use of a single p-value for final thresholding of the results [1]. In group fMRI studies, the need for compensation of cross-subjects misregistrations clearly justifies the smoothing. However, at the individual level, where neat delineations of the activated areas are of interest, the use of Gaussian smoothing as a pre-processing step is more questionable. In this paper, we propose to study the ability of an a contrario approach, recently adapted for basal perfusion abnormalities detection [2], to correctly detect areas of functional activity [42].

6.4.3. Robust perfusion maps in Arterial Spin Labeling by means of M-estimators

Participants: Camille Maumet, Pierre Maurel, Jean-Christophe Ferré, Christian Barillot.

Non-invasive measurement of Cerebral Blood Flow (CBF) is now feasible thanks to the introduction of Arterial Spin Labeling (ASL) Magnetic Resonance Imaging (MRI) techniques. To date, due to the low signal-to-noise ratio of ASL, a single acquisition (pair of control/label scans) is not sufficient to estimate perfusion reliably. Instead, the acquisition is usually repeated several times and the perfusion information is calculated by averaging across the repetitions. However, due to its zero breakdown point, the sample mean is very sensitive to outliers. In this paper, we propose to compute ASL CBF maps using Huber's M-estimator, a robust statistical function that is not overly impacted by outliers. This method is compared to an empirical approach, introduced in [1], based on z-score thresholding [43].

6.4.4. Quantifying CBF from Aterial Spin Labeling via Diverse-TI: sampling diversity or repetitions ?

Participants: Lei Yu, Pierre Maurel, Christian Barillot.

Arterial Spin Labeling (ASL) is a noninvasive perfusion technique which allows the absolute quantification of Cerebral Blood Flow (CBF). The perfusion is obtained from the difference between images with and without magnetic spin labeling of the arterial blood and the captured signal is around 0.5-2% of the magnitude of the labeling images, so the noise is one of the main problems for further data analysis. Classical method, *Mono-TI*, for CBF quantification is averaging repetitions with only one Inversion Time (TI) - the time delay between labeling and acquisition to allow the labeled blood to arrive the imaging slice. It improves the robustness to noise, however, cannot compensate the variety of Arterial Arrival Time (AAT). In this paper, *Diverse-TI* is proposed to exploit different TI sampling instants (sampling diversity) to improve the robustness to variety of AAT and simultaneously average repetitions with each TI (sampling repetitions) to improve the robustness to noise. Generally, the sampling diversity is relatively small and can be considered as compressed measurements, thus the Compressive Matched Filter (CMF) enlightened from sparsity is exploited to directly reconstruct CBF and AAT directly from compressed measurements. Meanwhile, regarding the CBF quantification performance, the compromise between the sampling repetition and sampling diversity is discussed and the empirical protocol to determine the sampling diversity is proposed. Simulations are carried out to highlight our discussions. This is a joint work with Remi Gribonval (Panama Team) [56].

6.4.5. Peripheral angiography and neurovascular imaging

Participants: Hélène Raoult, Jean-Yves Gauvrit, Elise Bannier, Pierre Maurel, Clement Neyton, Christian Barillot, Jean-Christophe Ferré.

Vascular imaging contributions were performed on two different regions during the evaluation period: first on peripheral angiography, then on neurovascular imaging. Arteriography and MR angiography are routinely performed in patients presenting vascular pathologies. Yet, contrast agent injection is contraindicated in patients with renal insufficiency and the underlying risk of developing nephrogenic systemic fibrosis further encourages research on non-contrast enhanced MR angiography techniques (NCE MRA). In this context, we have been working on new MR sequences to reliably detect vascular abnormalities.

A first study [29] was published, where we assessed the feasibility and image quality of an improved non-gated carotid NATIVE TrueFISP NCE MRA sequence providing an extended field of view and a shorter acquisition time as compared to Time-of-Flight (TOF) imaging. A second study [48] was recently accepted for publication in Radiology on intracranial NCE MRA for arteriovenous malformation imaging with a high temporal resolution over 2 cardiac cycles. Combined with image post-processing, it allows improved depiction of venous drainage necessary to evaluate hemorrhagic risk and quantification. This ongoing work was just submitted.