RESEARCH CENTER
Rennes - Bretagne-Atlantique

FIELD

Activity Report 2015

Section New Results

Edition: 2016-03-21

ALF Project-Team

7. New Results

7.1. Processor Architecture

Participants: Pierre Michaud, Bharath Narasimha Swamy, Sylvain Collange, Erven Rohou, André Seznec, Arthur Perais, Surya Khizakanchery Natarajan, Sajith Kalathingal, Tao Sun, Andrea Mondelli, Aswinkumar Sridharan, Biswabandan Panda, Fernando Endo.

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 (maybe 1000's) simpler, more silicon and power effective cores.

In the DAL research project, https://team.inria.fr/alf/members/andre-seznec/defying-amdahls-law-dal/, 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.

7.1.1. Microarchitecture

7.1.1.1. Branch prediction

Participant: André Seznec.

This research was done in collaboration with Joshua San Miguel and Jorge Albericio from University of Toronto

The most efficient branch predictors proposed in academic literature exploit both global branch history and local branch history. However, local history branch predictor components introduce major design challenges, particularly for the management of speculative histories. Therefore, most effective hardware designs use only global history components and very limited forms of local histories such as a loop predictor. The wormhole (WH) branch predictor was recently introduced to exploit branch outcome correlation in multidimensional loops. For some branches encapsulated in a multidimensional loop, their outcomes are correlated with those of the same branch in neighbor iterations, but in the previous outer loop iteration. Unfortunately, the practical implementation of the WH predictor is even more challenging than the implementation of local history predictors.

In [36], we introduce practical predictor components to exploit this branch outcome correlation in multidimensional loops: the IMLI-based predictor components. The iteration index of the inner most loop in an application can be efficiently monitored at instruction fetch time using the Inner Most Loop Iteration (IMLI) counter. The outcomes of some branches are strongly correlated with the value of this IMLI counter. A single PC+IMLI counter indexed table, the IMLI-SIC table, added to a neural component of any recent predictor (TAGE-based or perceptron-inspired) captures this correlation. Moreover, using the IMLI counter, one can efficiently manage the very long local histories of branches that are targeted by the WH predictor. A second IMLI-based component, IMLI-OH, allows for tracking the same set of hard-to-predict branches as WH. Managing the speculative states of the IMLI-based predictor components is quite simple. Our experiments show that augmenting a state-of-the-art global history predictor with IMLI components outperforms previous state-of-the-art academic predictors leveraging local and global history at much lower hardware complexity (i.e., smaller storage budget , smaller number of tables and simpler management of speculative states).

7.1.1.2. 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.

At a first step, we showed that all predictors are amenable to very high accuracy at the cost of some loss on prediction coverage [7]. 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 allows to leverage the possibility of validating predictions at commit to introduce a new microarchitecture, EOLE [19]. 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.

However complexity in the value predictor infrastructure itself is also problematic. First, multiple predictions must be generated each cycle, but multi-ported structures should be avoided. Second, the predictor should be small enough to be considered for implementation, yet coverage must remain high enough to increase performance. In [32], to address these remaining concerns, we first propose a block-based value prediction scheme mimicking current instruction fetch mechanisms, BeBoP. It associates the predicted values with a fetch block rather than distinct instructions. Second, to remedy the storage issue, we present the Differential VTAGE predictor. This new tightly coupled hybrid predictor covers instructions predictable by both VTAGE and Stride-based value predictors, and its hardware cost and complexity can be made similar to those of a modern branch predictor. Third, we show that block-based value prediction allows to implement the checkpointing mechanism needed to provide D-VTAGE with last computed/predicted values at moderate cost. Overall, we establish that EOLE with a 32.8KB block-based D-VTAGE predictor and a 4-issue OoO engine can significantly outperform a baseline 6-issue superscalar processor, by up to 62.2 % and 11.2 % on average (gmean), on our benchmark set.

The overall study on value prediction is presented in Arthur Perais's PhD [14].

7.1.1.3. Cost-Effective Speculative Scheduling in High Performance Processors

Participants: André Seznec, Arthur Perais, Pierre Michaud.

This study was done in collaboration with Andreas Sembrant and Erik Hagersten from Upsala University

To maximize performance, out-of-order execution processors sometimes issue instructions without having the guarantee that operands will be available in time; e.g. loads are typically assumed to hit in the L1 cache and dependent instructions are issued assuming a L1 hit. This form of speculation ?that we refer to as speculative scheduling? has been used for two decades in real processors, but has received little attention from the research community. In particular, as pipeline depth grows and the distance between the Issue and the Execute stages increases, it becomes critical to issue dependents on variable-latency instructions as soon as possible, rather than to wait for the actual cycle at which the result becomes available. Unfortunately, due to the uncertain nature of speculative scheduling, the scheduler may wrongly issue an instruction that will not have its source(s) on the bypass network when it reaches the Execute stage. Therefore, this instruction must be canceled and replayed, which can potentially impair performance and increase energy consumption.

In [31] we focus on ways to reduce the number of replays that are agnostic of the replay scheme. First, we propose an easily implementable, low-cost solution to reduce the number of replays caused by L1 bank conflicts. Schedule Shifting always assumes that, given a dual-load issue capacity, the second load issued in a given cycle will be delayed because of a bank conflict. Its dependents are thus always issued with a corresponding delay. Second, we also improve on existing L1 hit/miss prediction schemes by taking into account instruction criticality. That is, for some criterion of criticality and for loads whose hit/miss behavior is hard to predict, we show that it is more cost-effective to stall dependents if the load is not predicted critical. In total, in our experiments assuming a 4-cycle issue-to-execute delay, we found that the vast majority of instructions replays due to L1 data cache banks conflicts and L1 hit mispredictions can be avoided, thus leading to a 3.4% performance gain and a 13.4% decrease in the number of issued instructions, over a baseline speculative scheduling scheme.

7.1.1.4. Criticality-aware Resource Allocation in OOO Processors

Participants: André Seznec, Arthur Perais, Pierre Michaud.

This study was done in collaboration with Andreas Sembrant, Erik Hagersten, David Black-Schaffer and Trevor Carlson from Upsala University.

Modern processors employ large structures (IQ, LSQ, register file, etc.) to expose instruction-level parallelism (ILP) and memory-level parallelism (MLP). These resources are typically allocated to instructions in program order. This wastes resources by allocating resources to instructions that are not yet ready to be executed and by eagerly allocating resources to instructions that are not part of the application's critical path. In [35], we explore the possibility of allocating pipeline resources only when needed to expose MLP, and thereby enabling a processor design with significantly smaller structures, without sacrificing performance. First we identify the classes of instructions that should not reserve resources in program order and evaluate the potential performance gains we could achieve by delaying their allocations. We then use this information to "park" such instructions in a simpler, and therefore more efficient, Long Term Parking (LTP) structure. The LTP stores instructions until they are ready to execute, without allocating pipeline resources, and thereby keeps the pipeline available for instructions that can generate further MLP. LTP can accurately and rapidly identify which instructions to park, park them before they execute, wake them when needed to preserve performance, and do so using a simple queue instead of a complex IQ. We show that even a very simple queue-based LTP design allows us to significantly reduce IQ ($64 \rightarrow 32$) and register file ($128 \rightarrow 96$) sizes while retaining MLP performance and improving energy efficiency.

7.1.1.5. Efficient Execution on Guarded Instruction Sets

Participant: 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 and high confidence only guard prediction. 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 [8].

This study was accepted to ACM Transactions on Architecture and Compiler Optimizations (Dec. 2014) and presented at the HIPEAC conference in January 2015.

7.1.1.6. Clustered microarchitecture

Participants: Andrea Mondelli, Pierre Michaud, André Seznec.

In the last 10 years, the clock frequency of high-end superscalar processors did not increase significantly. Performance keeps being increased mainly by integrating more cores on the same chip and by introducing new instruction set extensions. However, this benefits only to some applications and requires rewriting and/or recompiling these applications. A more general way to increase performance is to increase the IPC, the number of instructions executed per cycle.

In [18], we argue that some of the benefits of technology scaling should be used to increase the IPC of future superscalar cores. Starting from microarchitecture parameters similar to recent commercial high-end cores, we show that an effective way to increase the IPC is to increase the issue width. But this must be done without impacting the clock cycle. We propose to combine two known techniques: clustering and register write specialization. The objective of past work on clustered microarchitecture was to allow a higher clock frequency while minimizing the IPC loss. This led researchers to consider narrow-issue clusters. Our objective, instead, is to increase the IPC without impacting the clock cycle, which means wide-issue clusters. We show that, on a wide-issue dual cluster, a very simple steering policy that sends 64 consecutive instructions to the same cluster, the next 64 instructions to the other cluster, and so on, permits tolerating an inter-cluster delay of several cycles. We also propose a method for decreasing the energy cost of sending results of one cluster to the other cluster.

7.1.1.7. Adaptive Intelligent Memory Systems

Participants: André Seznec, Aswinkumar Sridharan.

Multi-core processors employ shared Last Level Caches (LLC). This trend will continue in the future with large multi-core processors (16 cores and beyond) as well. At the same time, the associativity of this LLC tends to remain in the order of sixteen. Consequently, with large multicore processors, the number of cores that share the LLC becomes larger than the associativity of the cache itself. LLC management policies have been extensively studied for small scale multi-cores (4 to 8 cores) and associativity degree in the 16 range. However, the impact of LLC management on large multi-cores is essentially unknown, in particular when the associativity degree is smaller than the number of cores.

In [43], we introduce Adaptive Discrete and deprioritized Application PrioriTization (ADAPT), an LLC management policy addressing the large multi-cores where the LLC associativity degree is smaller than the number of cores. ADAPT builds on the use of the Footprint-number metric. Footprint-number is defined as the number of unique accesses (block addresses) that an application generates to a cache set in an interval of time. We propose a monitoring mechanism that dynamically samples cache sets to estimate the Footprint-number of applications and classifies them into discrete (distinct and more than two) priority buckets. The cache replacement policy leverages this classification and assigns priorities to cache lines of applications during cache replacement operations. Footprint-number is computed periodically to account the dynamic changes in applications behavior. We further find that de- prioritizing certain applications during cache replacement is beneficial to the overall performance. We evaluate our proposal on 16, 20 and 24-core multi-programmed workloads and discuss other aspects in detail.

[43] has been accepted for publication at the IPDPS 2016 conference.

7.1.1.8. Hardware data prefetching

Participant: Pierre Michaud.

Hardware prefetching is an important feature of modern high-performance processors. When an application's working set is too large to fit in on-chip caches, disabling hardware prefetchers may result in severe performance reduction. We propose a new hardware data prefetcher, the Best-Offset (BO) prefetcher. The BO prefetcher is an offset prefetcher using a new method for selecting the best prefetch offset taking into account prefetch timeliness. The hardware required for implementing the BO prefetcher is very simple. The BO prefetcher won the last Data Prefetching Championship [27].

A paper describing and studying the BO prefetcher has been accepted for publication at the HPCA 2016 conference.

7.1.1.9. Prediction-based superpage-friendly TLB designs

Participant: André Seznec.

This research was done in collaboration with Misel-Myrto Papadopoulou, Xin Tong and Andreas Moshovos from University of Toronto

In [30], we demonstrate that a set of commercial and scale-out applications exhibit significant use of superpages and thus suffer from the fixed and small superpage TLB structures of some modern core designs. Other processors better cope with superpages at the expense of using power-hungry and slow fully-associative TLBs. We consider alternate designs that allow all pages to freely share a single, power-efficient and fast set-associative TLB. We propose a prediction-guided multi-grain TLB design that uses a superpage prediction mechanism to avoid multiple lookups in the common case. In addition, we evaluate the previously proposed skewed TLB which builds on principles similar to those used in skewed associative caches . We enhance the original skewed TLB design by using page size prediction to increase its effective associativity. Our prediction-based multi-grain TLB design delivers more hits and is more power efficient than existing alternatives. The predictor uses a 32-byte prediction table indexed by base register values.

7.1.2. Microarchitecture Performance Modeling

7.1.2.1. Symbiotic scheduling on SMT cores and symmetric multicores

Participant: Pierre Michaud.

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

When several independent tasks execute concurrently on a simultaneous multithreaded (SMT) core or on a multicore, they share hardware resources. Hence the execution rate of a task is influenced by the other tasks running at the same time. Based on this observation, Snavely and Tullsen proposed *symbiotic* scheduling, i.e., the idea that performance can be increased by co-scheduling tasks that do not stress the same shared resources [63]. They claim that, when the number of concurrent tasks exceeds the number of logical cores, symbiotic scheduling increases performance substantially. A more recent study by Eyerman and Eeckhout reached similar conclusions [54].

We have revisited symbiotic scheduling for SMT cores and symmetric multicores [22], and we obtained very modest throughput gains, which seemingly contradicts the above mentioned studies. We analyzed the reasons for this discrepancy and found that previous studies did not measure throughput but average response time. Response time reductions can be magnified by setting the job arrival rate very close to the maximum throughput, which turns a tiny throughput increase into a large response time reduction. Also, the proposed scheduling policies are approximately equivalent to scheduling the shortest jobs first, which mechanically reduces the average response time independently of any symbiosis effect.

We identified three typical situations where symbiotic scheduling yields little to no throughput gain: (1) most of the time is spent executing a single type of job, or (2) jobs' execution rates barely depend on which other jobs are running concurrently, or (3) jobs' execution rates are proportional to the fraction they get of a certain shared resource (e.g., instruction decode bandwidth in an SMT core). In our experiments, most workloads were close to one of the three situations above.

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

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

Estimating the potential performance of parallel applications on the yet-to-be-designed future many cores is very speculative. The simple models proposed by Amdahl's law (fixed input problem size) or Gustafson's law (fixed number of cores) do not completely capture the scaling behaviour of a multi-threaded (MT) application leading to over estimation of performance in the many-core era. On the other hand, modeling many-core by simulation is too slow to study the applications performance. In [28], [13], we propose a more refined but still tractable, high level empirical performance model for multi-threaded applications, the Serial/Parallel Scaling (SPS) Model to study the scalability and performance of application in many-core era. SPS model learns the application behavior on a given architecture and provides realistic estimates of the performance in future many-cores. Considering both input problem size and the number of cores in modeling, SPS model can help in making high level decisions on the design choice of future many-core applications and architecture. We validate the model on the Many-Integrated Cores (MIC) xeon-phi with 240 logical cores.

7.1.2.3. Optimal cache replacement

Participant: Pierre Michaud.

This research was done in collaboration with Mun-Kyu Lee, Jeong Seop Sim and DaeHun Nyang from Inha University.

The replacement policy for a cache is the algorithm, implemented in hardware, selecting a block to evict for making room for an incoming block. This research topic has been revitalized in recent years. The MIN replacement policy, which evicts the block referenced furthest in the future, was introduced by Belady [51] and was later shown to be optimal by Mattson et al. [60]. The MIN policy is an offline policy that cannot be implemented in real processors, as it needs the knowledge of future memory accesses. Still, a possible way to improve online replacement policies would be to emulate the MIN policy, trying to use past references to predict future ones. However, the MIN policy is not intuitive, and Mattson et al.'s proof of optimality is quite involved. We believe that new intuition about the MIN policy will help microarchitects improve cache replacement policies. As a first step toward this goal, we produced a new, intuitive proof of optimality of the MIN policy [17].

7.1.3. Hardware/Software Approaches

7.1.3.1. Helper threads

Participants: Bharath Narasimha Swamy, André Seznec.

Heterogeneous Many Cores (HMC) architectures that mix many simple/small cores with a few complex/large cores are emerging as a design alternative that can provide both fast sequential performance for single threaded workloads and power-efficient execution for throughput oriented parallel workloads. The availability of many small cores in a HMC presents an opportunity to utilize them as low-power helper cores to accelerate memory-intensive sequential programs mapped to a large core. However, the latency overhead of accessing small cores in a loosely coupled system limits their utility as helper cores. Also, it is not clear if small cores can execute helper threads sufficiently in advance to benefit applications running on a larger, much powerful, core.

In [12] we present a hardware/software framework called core-tethering to support efficient helper threading on heterogeneous many-cores. Core-tethering provides a co-processor like interface to the small cores that (a) enables a large core to directly initiate and control helper execution on the helper core and (b) allows efficient transfer of execution context between the cores, thereby reducing the performance overhead of accessing small cores for helper execution. Our evaluation on a set of memory intensive programs chosen from the standard benchmark suites show that, helper threads using moderately sized small cores can significantly accelerate a larger core compared to using a hardware prefetcher alone. We also find that a small core provides a good trade-off against using an equivalent large core to run helper threads in a HMC.

In summary, despite the latency overheads of accessing prefetched cache lines from the shared L3 cache, helper thread based prefetching on small cores looks as a promising way to improve single thread performance on memory intensive workloads in HMC architectures.

This research was partially done in collaboration with Alain Ketterlin from the Inria Camus project-team in Strasbourg.

7.1.3.2. 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. Folklore attributes a significant penalty to this branch, due to its high misprediction rate. In [34], we revisit this assumption, considering state-of-the-art branch predictors and the three most recent Intel processor generations on current interpreters. Using both hardware counters on Haswell, the latest Intel processor generation, and simulation of the ITTAGE predictor [10], we show that the accuracy of indirect branch prediction is no longer critical for interpreters. We further compare the characteristics of these interpreters and analyze why the indirect branch is less important than before.

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

Participants: Sylvain Collange, André Seznec, Sajith Kalathingal.

Threads of Single-Program Multiple-Data (SPMD) applications often exhibit very similar control flows, i.e. they execute the same instructions on different data. In [42] we propose the Dynamic Inter-Thread Vectorization Architecture (DITVA) to leverage this implicit Data Level Parallelism on SPMD applications to create dynamic vector instructions at runtime. DITVA extends an in-order SMT processor with SIMD units with an inter-thread vectorization execution mode. In this mode, identical instructions of several threads running in lockstep are aggregated into a single SIMD instruction. DITVA leverages existing SIMD units and maintains binary compatibility with existing CPU architectures. To balance TLP and DLP, threads are statically grouped into fixed-size warps, inside which threads run in lockstep. At instruction fetch time, if the instruction streams of several threads within a warp are synchronized, then DITVA aggregates the instructions of the threads as dynamic vectors. To maximize vectorization opportunities, we use resource sharing arbitration policies that favor thread synchronization within warps. The policies do not require any compiler hints or modified algorithms for the existing SPMD applications and allow to run unmodified CPU binaries. A dynamic vector instruction is executed as a single unit. This allows to execute m identical instructions from m different threads on m parallel execution lanes while activating the I-fetch, the decode, and the overall pipeline control only once.

Our evaluation on the SPMD applications from the PARSEC and SPLASH benchmarks shows that a 4-warp 4-lane 4-issue DITVA architecture with a realistic bank-interleaved cache achieves 44% higher performance than a 4-thread 4-issue SMT architecture with AVX instructions while fetching and issuing 40 % fewer instrructions, achieving an overall 22% energy reduction.

7.2. Compiler, vectorization, interpretation

Participants: Erven Rohou, Emmanuel Riou, Bharath Narasimha Swamy, Arjun Suresh, André Seznec, Nabil Hallou, Sylvain Collange.

7.2.1. Improving sequential performance through memoization

Participants: Erven Rohou, Emmanuel Riou, Bharath Narasimha Swamy, André Seznec, Arjun Suresh.

Many applications perform repetitive computations, even when properly programmed and optimized. Performance can be improved by caching results of pure functions, and retrieving them instead of recomputing a result (a technique called memoization).

We propose [20] a simple technique for enabling software memoization of any dynamically linked pure function and we illustrate our framework using a set of computationally expensive pure functions – the transcendental functions.

Our technique does not need the availability of source code and thus can be applied even to commercial applications as well as applications with legacy codes. As far as users are concerned, enabling memoization is as simple as setting an environment variable.

Our framework does not make any specific assumptions about the underlying architecture or compiler tool-chains, and can work with a variety of current architectures.

We present experimental results for x86-64 platform using both gcc and icc compiler tool-chains, and for ARM cortex-A9 platform using gcc. Our experiments include a mix of real world programs and standard benchmark suites: SPEC and Splash2x. On standard benchmark applications that extensively call the transcendental functions we report memoization benefits of upto 16 %, while much higher gains were realized for programs that call the expensive Bessel functions. Memoization was also able to regain a performance loss of 76 % in *bwaves* due to a known performance bug in the gcc libm implementation of *pow* function.

This work has been published in ACM TACO 2015 [20] and accepted for presentation at the International Conference HiPEAC 2016.

7.2.2. 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 propose [24] 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.

This work was presented in January 2015 at the International Workshop on Dynamic Compilation Everywhere (DCE-2015) [24].

7.2.3. Dynamic Binary Re-vectorization

Participants: Erven Rohou, Nabil Hallou, Emmanuel Riou.

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

Applications are often under-optimized for the hardware on which they run. Several reasons contribute to this unsatisfying situation, including the use of legacy code, commercial code distributed in binary form, or deployment on compute farms. In fact, backward compatibility of instruction sets guarantees only the functionality, not the best exploitation of the hardware. In particular SIMD instruction sets are always evolving.

We proposed [23] a runtime re-vectorization platform that dynamically adapts applications to execution hardware. The platform is built on top of Padrone. Programs distributed in binary forms are re-vectorized at runtime for the underlying execution hardware. Focusing on the x86 SIMD extensions, we are able to automatically convert loops vectorized for SSE into the more recent and powerful AVX. A lightweight mechanism leverages the sophisticated technology put in a static vectorizer and adjusts, at minimal cost, the width of vectorized loops. We achieve speedups in line with a native compiler targeting AVX. Our re-vectorizer is implemented inside a dynamic optimization platform; its usage is completely transparent to the user and requires neither access to source code nor rewriting binaries.

7.2.4. Dynamic Parallelization of Binary Executables

Participants: Erven Rohou, Nabil Hallou, Emmanuel Riou.

We address runtime automatic parallelization of binary executables, assuming no previous knowledge on the executable code. The Padrone platform is used to identify candidate functions and loops. Then we disassemble the loops and convert them to the intermediate representation of the LLVM compiler (thanks to the external tool McSema). This allows us to leverage the power of the polyhedral model for auto-parallelizing loops. Once optimized, new native code is generated just-in-time in the address space of the target process.

Our approach enables user transparent auto-parallelization of legacy and/or commercial applications with auto-parallelization.

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

7.2.5. Hardware Accelerated JIT Compilation for Embedded VLIW Processors

Participant: Erven Rohou.

Just-in-time (JIT) compilation is widely used in current embedded systems (mainly because of Java Virtual Machine). When targeting Very Long Instruction Word (VLIW) processors, JIT compilation back-ends grow more complex because of the instruction scheduling phase. This tends to reduce the benefits of JIT compilation for such systems. We propose a hybrid JIT compiler where JIT management is handled in software and the back-end is performed by specialized hardware. Experimental studies show that this approach leads to a compilation up to 15 times faster and 18 times more energy efficient than a pure software compilation.

This work is done in collaboration with the CAIRN team (Steven Derrien and Simon Rokicki).

7.2.6. Performance Assessment of Sequential Code

Participant: Erven Rohou.

The advent of multicore and manycore processors, including GPUs, in the customer market encouraged developers to focus on extraction of parallelism. While it is certainly true that parallelism can deliver performance boosts, parallelization is also a very complex and error-prone task, and many applications are still dominated by sequential sections. Micro-architectures have become extremely complex, and they usually do a very good job at executing fast a given sequence of instructions. When they occasionally fail, however, the penalty is severe. Pathological behaviors often have their roots in very low-level details of the micro-architecture, hardly available to the programmer. In [33], we argue that the impact of these low-level features on performance has been overlooked, often relegated to experts. We show that a few metrics can be easily defined to help assess the overall performance of an application, and quickly diagnose a problem. Finally, we illustrate our claim with a simple prototype, along with use cases.

7.2.7. Compilers for emerging throughput architectures

Participant: Sylvain Collange.

This work is done in collaboration with Douglas de Couto and Fernando Pereira from UFMG.

The increasing popularity of Graphics Processing Units (GPUs) has brought renewed attention to old problems related to the Single Instruction, Multiple Data execution model. One of these problems is the reconvergence of divergent threads. A divergence happens at a conditional branch when different threads disagree on the path to follow upon reaching this split point. Divergences may impose a heavy burden on the performance of parallel programs.

We have proposed a compiler-level optimization to mitigate the performance loss due to branch divergence on GPUs. This optimization consists in merging function call sites located at different paths that sprout from the same branch. We show that our optimization adds negligible overhead on the compiler. When not applicable, it does not slow down programs and it accelerates substantially those in which it is applicable. As an example, we have been able to speed up the well known SPLASH Fast Fourier Transform benchmark by 11 %.

7.2.8. Deterministic floating-point primitives for high-performance computing

Participant: Sylvain Collange.

This work is done in collaboration with David Defour (UPVD), Stef Graillat and Roman Iakymchuk (LIP6).

Parallel algorithms such as reduction are ubiquitous in parallel programming, and especially high-performance computing. Although these algorithms rely on associativity, they are used on floating-point data, on which operations are not associative. As a result, computations become non-deterministic, and the result may change according to static and dynamic parameters such as machine configuration or task scheduling.

We introduced a solution to compute deterministic sums of floating-point numbers efficiently and with the best possible accuracy. A multi-level algorithm incorporating a filtering stage that uses fast vectorized floating-point expansions and an accumulation stage based on superaccumulators in a high-radix carry-save representation guarantees accuracy to the last bit even on degenerate cases while maintaining high performance in the common cases [16]. Leveraging these algorithms, we build a reproducible BLAS library [49] and extend the approach to triangular solvers [25].

7.3. WCET estimation and optimization

Participants: Hanbing Li, Isabelle Puaut, Erven Rohou, Damien Hardy, Viet Anh Nguyen, Benjamin Rouxel.

7.3.1. WCET estimation for architectures with faulty caches

Participants: Damien Hardy, Isabelle Puaut.

This is joint work with Yannakis Sazeides from University of Cyprus

Fine-grained disabling and reconfiguration of hardware elements (functional units, cache blocks) will become economically necessary to recover from permanent failures, whose rate is expected to increase dramatically in the near future. This fine-grained disabling will lead to degraded performance as compared to a fault-free execution.

Until recently, all static worst-case execution time (WCET) estimation methods were assuming fault-free processors, resulting in unsafe estimates in the presence of faults. The first static WCET estimation technique dealing with the presence of permanent faults in instruction caches was proposed in [4]. This study probabilistically quantified the impact of permanent faults on WCET estimates. It demonstrated that the probabilistic WCET (pWCET) estimates of tasks increase rapidly with the probability of faults as compared to fault-free WCET estimates.

New results show that very simple reliability mechanisms allow mitigating the impact of faulty cache blocks on pWCETs. Two mechanisms, that make part of the cache resilient to faults are analyzed. Experiments show that the gain in pWCET for these two mechanisms are on average 48% and 40% as compared to an architecture with no reliability mechanism.

This work will appear at DATE 2016.

7.3.2. Speeding up Static Probabilistic Timing Analysis

Participants: Damien Hardy, Isabelle Puaut.

This is joint work with Suzana Milutinovic, Jaume Abella, Eduardo Quinones and Francisco J. Cazorla from Barcelona Supercomputing Center.

Probabilistic Timing Analysis (PTA) has emerged recently to derive trustworthy and tight WCET estimates. For its static variant, called SPTA, we identify one of the main elements that jeopardizes its scalability to real-size programs: its high computation time cost. This SPTA's high computational costs are due to convolution, a mathematical operator used by SPTA and also deployed in many domains including signal and image processing.

In [40], we show how convolution is applied in SPTA, and qualitatively and quantitatively evaluate optimizations developed in other domains to reduce convolution time cost when applied to SPTA, and SPTA-specific optimizations. We show that SPTA-specific optimizations provide larger execution time reductions than generic cores.

7.3.3. 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 through 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 our recent work, we have proposed a framework to systematically transform flow information from source code to machine code. The framework [11] defines a set of formulas to transform flow information for standard compiler optimizations. Transforming the flow information is done within the compiler, in parallel with transforming the code. There thus is no guessing what flow information have become, it is transformed along with the code.

Our most recent results in this framework were to add support for vectorization [26]. We implemented our approach in the LLVM compiler. In addition, we show through measurements on single-path programs that vectorization improves not only average-case performance but also WCETs. The WCET improvement ratio ranges from 1.18x to 1.41x depending on the target architecture on a benchmark suite designed for vectorizing compilers (TSVC).

This work is part of a more general traceability framework, designed and implemented within the ANR W-SEPT project and described in paper [21]. In this paper, we introduce a complete semantic-aware WCET estimation workflow. We introduce some program analysis to find infeasible paths: they can be performed at design, C or binary level, and may take into account information provided by the user. We design an annotation-aware compilation process that enables to trace the infeasible path properties through the program transformations performed by the compilers. Finally, we adapt the WCET estimation tool to take into account the kind of annotations produced by the workflow.

7.3.4. WCET estimation for many core processors

Participants: Viet Anh Nguyen, Damien Hardy, Isabelle Puaut.

This research is part of the PIA Capacités project.

The overall goal of this research is to defined WCET estimation methods for parallel applications running on many-core architectures, such as the Kalray MPPA machine.

Some approaches to reach this goal have been proposed, but they assume the mapping of parallel applications on cores already done. Unfortunately, on architectures with caches, task mapping requires a priori known WCETs for tasks, which in turn requires knowing task mapping (i.e., co-located tasks, co-running tasks) to have tight WCET bounds. Therefore, scheduling parallel applications and estimating their WCET introduce a chicken and egg situation.

In [41], we address this issue by developing an optimal integer linear programming formulation for solving the scheduling problem, whose objective is to minimize the WCET of a parallel application. Our proposed static partitioned non-preemptive mapping strategy addresses the effect of local caches to tighten the estimated WCET of the parallel application. We report preliminary results obtained on synthetic parallel applications.

CAIRN Project-Team

7. New Results

7.1. Reconfigurable Architecture Design

7.1.1. Design Flow and Run-Time Management for Compressed FPGA Configurations

Participants: Olivier Sentieys, Christophe Huriaux.

Almost since the creation of the first SRAM-based FPGAs there has been a desire to explore the benefits of partially reconfiguring a portion of an FPGA at run-time while the remainder of design functionality continues to operate uninterrupted. Currently, the use of partial reconfiguration imposes significant limitations on the FPGA design: reconfiguration regions must be constrained to certain shapes and sizes and, in many cases, bitstreams must be precompiled before application execution depending on the precise region of the placement in the fabric. We developed an FPGA architecture that allows for seamless translation of partially-reconfigurable regions, even if the relative placement of fixed-function blocks within the region is changed.

In [42] we proposed a design flow for generating compressed configuration bit-streams abstracted from their final position on the logic fabric. Those configurations can then be decoded and finalized in real-time and at run-time by a dedicated reconfiguration controller to be placed at a given physical location. The VTR framework has been expanded to include bit-stream generation features. A bit-stream format is proposed to take part of our approach and the associated decoding architecture was designed. We analyzed the compression induced by our coding method and proved that compression ratios of at least $2.5\times$ can be achieved on the 20 largest MCNC benchmarks. The introduction of clustering which aggregates multiple routing resources together showed compression ratio up to a factor of $10\times$, at the cost of a more complex decoding step at runtime. The VBS approach can provide increased online relocation capabilities using a decoding algorithm capable of decoding the VBS on-the-fly during the task migration.

7.1.2. Run-Time Approximation under Performance Constraints in OFDM Wireless Receivers

Participants: Olivier Sentieys, Fernado Cladera.

Mobile wireless channels are characterized by time-varying multipath propagation, noise, and interference effects. To cope with these rapid variations of channel parameters, wireless receivers are designed with a significant performance margin to be able to reach a given link quality (BER - Bit Error Rate), even for the worst-case channel conditions. Indeed, one of the steps during the design phase is the choice of the architecture bit-width, and the smallest wordlength that ensures the correct behaviour of the receiver is usually chosen. In [39], an adaptive precision OFDM receiver is proposed. 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 a FPGA and a processor using soft SIMD. Results show that up to 62% of the dynamic energy consumption can be saved using this adaptive technique. The algorithms proposed for the low complexity selector used to choose the processing word-length at run time, without modifying the standard OFDM frame, are detailed in [38].

7.1.3. Optical Interconnections for 3D Multiprocessor Architectures

Participants: Jiating Luo, Pham Van Dung, Cédric Killian, Daniel Chillet, Olivier Sentieys.

To address the issue of interconnection bottleneck in multiprocessor on a single chip, we study how an Optical Network-on-Chip (ONoC) can leverage 3D technology by stacking a specific photonics die. The objectives of this study target: i) the definition of a generic architecture including both electrical and optical components, ii) the interface between electrical and optical domains, iii) the definition of strategies (communication protocol) to manage this communication medium, and iv) new techniques to manage and reduce the power consumption of optical communications. The first point is required to ensure that electrical and optical components can be used together to define a global architecture. Indeed, optical components are generally larger than electrical components, so a trade-off must be found between the size of optical and electrical parts. For example, if the need in terms of communications is high, several waveguides and wavelengths must be necessary, and can lead to an optical area larger than the footprint of a single processor. In this case, a solution is to connect (through the optical NoC) clusters of processors rather than each single processor. For the second point, we study how the interface can be designed to take applications needs into account. From the different possible interface designs, we extract a high-level performance model of optical communications from losses induced by all optical components to efficiently manage Laser parameters. Then, the third point concerns the definition of high-level mechanisms which can handle the allocation of the communication medium for each data transfer between tasks. This part consists in defining the protocol of wavelength allocation. Indeed, the optical wavelengths are a shared resource between all the electrical computing clusters and are allocated at run time according to application needs and quality of service. The last point concerns the definition of techniques allowing to reduce the power consumption of on-chip optical communications. The power of each Laser can be dynamically tuned in the optical/electrical interface at run time for a given targeted bit-error-rate. Due to the relatively high power consumption of such integrated Laser, we study how to define adequate policies able to adapt the laser power to the signal losses.

In [44], we proposed a wavelength reservation protocol handled by an Optical Network Interface (ONI) Manager for reconfigurable ONoC based on shared waveguide. It allows to efficiently allocate, at runtime, the optical communication channels for a manycore architecture. We described the ONI manager architecture and reservation protocol. Synthesis results in a 28nm FDSOI technology demonstrated that our interface can support a clock frequency up to 550 MHz with 6 wavelengths managed. From these results, we can be optimistic about the scaling of the ONoC and its capacity to manage a large number of processors and more wavelengths.

In [55], we explored the trade-off among channel bandwidth alternatives, performance, area and power. We showed that the channel size has a strong impact on the system performance and cost. We employed synthetic and real application traffic executed on the GEM5 simulator. As a result, we show that different channel bandwidths can improve the execution time of an application up to 75%, while including low area and power penalties.

7.1.4. Arithmetic Operators for Cryptography and Fault-Tolerance

Participants: Arnaud Tisserand, Emmanuel Casseau, Nicolas Veyrat-Charvillon, Karim Bigou, Franck Bucheron, Jérémie Métairie, Gabriel Gallin.

Arithmetic Operators for Fast and Secure Cryptography.

Our paper [36], presented at CHES, describes a new RNS modular multiplication algorithm for efficient implementations of ECC over GF(p). Thanks to the proposition of RNS-friendly Mersenne-like primes, the proposed RNS algorithm requires 2 times less moduli than the state-of-art ones, leading to 4 times less precomputations and about 2 times less operations. FPGA implementations of our algorithm are presented, with area reduced up to 46 %, for a time overhead less than 10 %. Other RNS algorithms and implementations have been presented at RAIM [66].

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). Ensuring uniform computation profiles is an efficient protection against some side channel attacks (SCA) in embedded systems. Typical ECC scalar multiplication methods use two point operations (addition and doubling) scheduled according to secret scalar digits. Euclidean addition chains (EAC) offer a natural SCA protection since only one point operation

is used. Computing short EACs is considered as a very costly operation and no hardware implementation has been reported yet. We designed an hardware recoding unit for short EACs which works concurrently to scalar multiplication. It has been integrated in an in-house ECC processor on various FPGAs. The implementation results show similar computation times compared to non-protected solutions, and faster ones compared to typical protected solutions (e. g. 18 % speed-up over 192 b Montgomery ladder). A paper [62] has been presented at Compas conference.

In a collaboration with University College Cork (Ireland), we worked on the design of secure multipliers for asymmetric cryptography using asynchronous circuits. A common paper has been published at ASYNC Conference [37]. In this paper, a specially adjusted Latch-less Asynchronous Charge Sharing Logic (LACSL) is developed to inherently defend such architecture against DPA attacks. The proposed logic provides input data independent low-power/energy consumption which is attributed to interleaved charge sharing stages with non-static elements involved in the data path. A 32-bit LACSL Montgomery Multiplier (case study) is extensively tested through HSPICE simulations and great consistency in power/energy consumption is achieved. The normalized energy deviation and normalized standard deviation are only 0.048 and 0.011, respectively. Compared with the original ACSL implementation, besides the impressive energy coherence, 42% energy saving is demonstrated plus that the leakage power is 3.5 times smaller. Furthermore, the scalability of the proposed multiplier is explored where 64- bit, 128-bit and 256-bit designs are implemented. Again, great energy consistency is found with the highest deviation being 0.5%.

In collaboration with D. Pamula, we worked on fast and secure finite field multipliers for GF(2^m) arithmetic, a paper has been presented at DSD conference [53]. It presents details on fast and secure GF(2^m) multipliers dedicated to elliptic curve cryptography applications. Presented design approach aims at high efficiency and security against side channel attacks of a hardware multi- plier. The security concern in the design process of a GF(2^m) multiplier is quite a novel concept. Basing on the results obtained in course of conducted research it is argued that, as well as efficiency of the multiplier impacts the efficiency of the cryptoprocessor, the security level of the multiplier impacts the security level of the whole cryptoprocessor. Thus the goal is to find a tradeoff, to compromise efficiency, in terms of speed and area, and security of the multiplier. We intend to secure the multiplier by masking the operation, either by uniformization or by randomization of the power consumption of the device during its work. The design methodology is half automated. The analyzed field sizes are the standard ones, which ensure that a cryptographic system is mathematically safe. The described architecture is based on principles of Mastrovito multiplication method. It is very flexible and enables to improve the resistance against side channel attacks without degrading the multiplier efficiency.

In a collaboration with G. Abozaid (EJUST University Egypt), we worked on the FPGA implementation of arithmetic operators for very large numbers (millions of bits) in fully homomorphic encryption (FHE) applications. A journal paper has been published in IEEE Embedded Systems Letters [18].

ECC Crypto-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. An ASIC version of the processor is under development and should be sent for fabrication in the beginning of 2016.

A. Tisserand has been invited speaker at the conference on elliptic curve cryptography (ECC): "Hardware Accelerators for ECC and HECC" [29].

Arithmetic Operators and Crypto-Processor for HECC.

In the HAH project, we study and prototype efficient arithmetic algorithms for hyperelliptic curve cryptography for hardware implementations (on FPGA circuits). We study new advanced arithmetic algorithms and representations of numbers for efficient and secure implementations of HECC in hardware. First results have been published in Compas conference [60] and RAIM workshop [68].

Arithmetic Operators for Fault Tolerance.

In the ARDyT and Reliasic projects, 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.

7.2. Compilation and Synthesis for Reconfigurable Platform

7.2.1. Adaptive dynamic compilation for low power embedded systems

Participants: Steven Derrien, Simon Rokicki.

Just-in-time (JIT) compilers have been introduced in the 1960s and became popular in the mid-1990s with the Java virtual machine. The use of JIT techniques for bytecode languages brings both portability and performance, making it an attractive solution for embedded systems, as evidenced by the Dalvik framework used by Android.

When targeting embedded systems, JIT compilation is even more challenging. First, because embedded systems are often based on architectures with an explicit use of Instruction- Level Parallelism (ILP), such as Very Long Instruction Word (VLIW) processors. Those architectures are highly dependent of the quality of the compilation, mainly because of the instruction scheduling phase performed by the compiler. The other challenge lies in the high constraints of the embedded system: the energy and execution time overhead due to the JIT compilation must be carefully kept under control. This is even more true if the JIT system is to be used in the context of a heterogeneous multi-core system with support dynamic task migration for heterogeneous ISA cores and/or support dynamically reconfigurable machines.

To address these challenges, we are currently studying how it is possible to take advantage of custom hardware to speed-up (and reduce the energy cost of) the JIT compilation stage. In this framework, basic optimizations and JIT management are performed in software, while the compilation back-end is implemented by means of specialized hardware. This back-end involves both instruction scheduling and register allocation, which are known to be the most time consuming stages of such a compiler. The first results are very encouraging, and we are finalizing an FPGA-based demonstration of the system.

7.2.2. Design Tools for Reconfigurable Video Coding

Participants: Emmanuel Casseau, Yaset Oliva.

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 design flow starts with the Dynamic Dataflow and CAL descriptions of an application and goes up to the deployment of the system onto the hardware platform. We also propose a framework to explore dynamic mapping algorithms for multiprocessors systems. Such an algorithm should be capable of computing a more efficient workload repartition based on the current configuration and performances of the system. The targeted platform is composed of several Processing Elements (PE). They follow a hierarchical organization: one PE plays the role of master and the others are slaves. The master assigns tasks (actors) to the slaves. The slaves execute the application tasks. The system has been implemented on a Zynq platform. The mapping is computed at runtime on the ARM processor while two clusters of 8 Microblazes each play the role of slaves. The DDR memory is split into two sections: one is reserved to the

Master and the other one is shared with the slaves. This later contains the actor's code. On the FPGA, the Microblazes are connected to private memories through the Local Memory Bus (LMB) that store the runtime copy. A common shared memory is used for the data exchanges between the processors. It contains the FIFOs for token exchanges between actors. The dynamic mapping algorithm aims at increasing data throughput. It starts by gathering the performance metrics of the system. It then identifies the processor with the highest workload. The algorithm evaluates the gain when moving the actor to one of the other processors. The migration is only valuable if the overhead of moving the actor is less that the gain. The actor that would lead to the highest gain is selected for migration. As a use case, we implement an MPEG-4 decoder algorithm onto a multi-core heterogeneous system deployed onto the Zynq platform from Xilinx [61] [69]. This work is done in collaboration with Lab-STICC Lorient.

7.2.3. High-Level Synthesis Based Rapid Prototyping of Software Radio Waveforms

Participants: Emmanuel Casseau, Mai Thanh Tran.

Software Defined Radio (SDR) is now becoming a ubiquitous concept to describe and implement Physical Layers (PHYs) of wireless systems. In this context, FPGA (Field Programmable Gate Array) technology is expected to play a key role. To this aim, leveraging the nascent High-Level Synthesis (HLS) tools, a design flow from high-level specifications to Register-Transfer Level (RTL) description can be thought to generate processing blocks that can be reconfigured at run-time. We thus propose a methodology for the implementation of run-time reconfiguration in the context of FPGA-based SDR. The design flow allows the exploration between dynamic partial reconfiguration and control signal based multi-mode design. This architectural tradeoff relies upon HLS and its associated design optimizations. We apply the methodology to the architectural exploration of a Fast Fourier Transform (FFT) for Long Term Evolution (LTE) standard as a use case.

7.2.4. Optimization of loop kernels using software and memory information

Participant: Angeliki Kritikakou.

Compilers optimize the compilation sub-problems one after the other, following an order which leads to less efficient solutions because the different sub-problems are independently optimized taking into account only a part of the information available in the algorithms and the architecture. In [19], we have presented an approach which applies loop transformations in order to increase the performance of loop kernels. The proposed approach focuses on reducing the L1, L2 data cache and main memory accesses and the addressing instructions. Our approach exploits the software information, such as the array subscript equations, and the memory architecture, such as the memory sizes. Then, it applies source-to-source transformations taking as input the C code of the loop kernels and producing a new C code which is compiled by the target compiler. We have applied our approach to five well-known loop kernels for both embedded processors and general purpose processors. From the obtained experimental results we observed speedup gains from 2 up to 18. [21] presents a new methodology for computing the Dense Matrix Vector Multiplication, for both embedded (processors without SIMD unit) and general purpose processors (single and multi-core processors with SIMD unit). The proposed methodology fully exploits the combination of the software (e.g., data reuse) and hardware parameters (e.g., data cache associativity) which are considered simultaneously giving a smaller search space and high-quality solutions. The proposed methodology produces a different schedule for different values of the (i) number of the levels of data cache; (ii) data cache sizes; (iii) data cache associativities; (iv) data cache and main memory latencies; (v) data array layout of the matrix and (vi) number of cores. With our experimental results we show that the proposed approach achieves increased performance than ATLAS state-of-the-art library with a speedup from 1.2 up to 1.45.

7.2.5. Leveraging Power Spectral Density for Scalable System-Level Accuracy Evaluation

Participants: Benjamin Barrois, Olivier Sentieys.

The choice of fixed-point word-lengths critically impacts the system performance by impacting the quality of computation, its energy, speed and area. Making a good choice of fixed-point word-length generally requires solving an NP-hard problem by exploring a vast search space. Therefore, the entire fixed-point refinement process becomes critically dependent on evaluating the effects of accuracy degradation. In [34], a novel technique for the system-level evaluation of fixed-point systems, which is more scalable and that renders better accuracy, was proposed. This technique makes use of the information hidden in the power-spectral density of quantization noises. It is shown to be very effective in systems consisting of more than one frequency sensitive components. Compared to state-of-the-art hierarchical methods that are agnostic to the quantization noise spectrum, we show that the proposed approach is $5\times$ to $500\times$ more accurate on some representative signal processing kernels.

7.3. Interaction between Algorithms and Architectures

7.3.1. Sensor-Aided Non-Intrusive Load Monitoring

Participants: Xuan-Chien Le, Olivier Sentieys.

Non-Intrusive Load Monitoring (NILM) plays an important role in energy management and energy reduction in buildings and homes. An NILM system does not need a large amount of deployed power meters to monitor the power usage of home devices. Instead, only one meter on the main power line is necessary to detect and identify the operating devices. There are many approaches to solve the problem of device determination in NILM. The features applied in low-frequency based approach essentially include the step-change (or edge) and the steady state. In [47] we introduced three algorithms to solve the $l1$-norm minimization problem in NILM and results on power measurements obtained from a real appliance deployment. With a small number of devices, the obtained precision varies from 75% to 99%, depending on the tolerance criterion to determine the steady state of a given device.

7.3.2. Posture and Gesture Recognition using Wireless Body Sensor Networks

Participants: Arnaud Carer, Alexis Aulery, Olivier Sentieys.

The BoWi project (Body Wold Interactions) aims at designing a Wireless Body Sensor Network (WBSN) for accurate Gesture and Body Movement estimation with extremely severe constraints in terms of footprint and power consumption. Advantages of such system mainly come from its possible use in indoor or outdoor environments without any additional equipment. The 3D geolocation approach will combine radio communication distance measurement and inertial sensors and it will also strongly benefit from cooperative techniques based on multiple observations and distributed computation. Different types of applications, as health care, activity monitoring and environment control, are considered and evaluated along with a human-machine interface expertise.

In [32] we presented three different use cases of WBSN for posture and gesture recognition developed by increasing demands in terms of accuracy: posture recognition, gesture recognition and motion capture. This work is based on a simulator designed to explore algorithmic solutions for posture and gesture identification. Simulation results were performed with a set of different algorithm and sensor proposals for three usages including a Principal Component Analysis (PCA) for posture classification. We show how sensor and algorithm can be carefully chosen according to application scenarios while minimising implementation complexity.

For applications based on predefined postures such as environment control and physical rehabilitation, we show in [31] that low cost and fully distributed solutions, that minimize radio communications, can be efficiently implemented. Considering that radio links provide distance information, we also demonstrate that the matrix of estimated inter-node distances offers complementary information that allows for the reduction of communication load. Our results are based on a simulator that can handle various measured input data, different algorithms and various noise models. Simulation results are useful and used for the development of real-life prototype.

7.3.3. Energy Harvesting and Power Management

Participants: Olivier Sentieys, Arnaud Carer, Trong-Nhan Le.

To design autonomous Wireless Sensor Networks (WSNs) with a theoretical infinite lifetime, energy harvesting (EH) techniques have been recently considered as promising approaches. Ambient sources can provide everlasting additional energy for WSN nodes and exclude their dependence on battery.

In [24], an efficient energy harvesting system which is compatible with various environmental sources, such as light, heat, or wind energy, was proposed. Our platform takes advantage of double-level capacitors not only to prolong system lifetime but also to enable robust booting from the exhausting energy of the system. Simulations and experiments show that our multiple-energy-sources converter (MESC) can achieve booting time in order of seconds. Although capacitors have virtual recharge cycles, they suffer higher leakage compared to rechargeable batteries. Increasing their size can decrease the system performance due to leakage energy. Therefore, an energy-neutral design framework providing a methodology to determine the minimum size of those storage devices satisfying energy-neutral operation (ENO) and maximizing system quality-of-service (QoS) in EH nodes, when using a given energy source, was also proposed. Experiments validating this framework are performed on a real WSN platform with both photovoltaic cells and thermal generators in an indoor environment. Moreover, simulations on OMNET++ showed that the energy storage optimized from our design framework is used up to 93.86%.

A Power Manager (PM) is usually embedded in EH wireless nodes to adapt the computation load by changing their wake-up interval according to the harvested energy. In order to prolong the network lifetime, the PM must ensure that every node satisfies the Energy Neutral Operation (ENO) condition. However, when a multi-hop network is considered, changing the wake-up interval regularly may cripple the synchronization among nodes and therefore, degrade the global system Quality of Service (QoS). In [25], a Wake-up Variation Reduction Power Manager (WVR-PM) was proposed to solve this issue. This PM is applied for wireless nodes powered by a periodic energy source (e.g. light energy in an office) over a constant cycle of 24 hours. Not only following the ENO condition, our power manager also reduces the wake-up interval variations of WSN nodes. Based on this PM, an energy-efficient protocol, named Synchronized Wake-up Interval MAC (SyWiM), was also proposed. OMNET++ simulation results using three different harvested profiles show that the data rate of a WSN node can be increased up to 65% and the latency reduced down to 57% compared to state-of-the-art PMs. Validations on a real WSN platform have also been performed and confirmed the efficiency of our approach.

7.3.4. Signal Processing for High-Rate Optical Communications

Participants: Trung-Hien Nguyen, Olivier Sentieys, Arnaud Carer.

Mary quadrature amplitude modulation (m-QAM) combined with coherent detection and digital signal processing (DSP) is a promising candidate for the implementation of next generation optical transmission systems. However, as the number of modulation levels increases, the sensitivity to system imperfections such as phase noise of the transmitter and the local oscillator lasers or fiber nonlinearities is exacerbated. Moreover, the amplitude and phase imbalances between the in-phase (I) and quadrature (Q) channels in the transmitter (Tx) and the front-end of the receiver (Rx), which is often referred to as IQ imbalance, is also troublesome if not compensated

In [52], we proposed a novel simple blind adaptive IQ imbalance compensation based on a decision-directed least-mean-square (DD-LMS) algorithm integrated to a modified butterfly FIR filter configuration. Since only 2 FIR filter coefficients-sets are used, instead of 4 in the conventional configuration, the time for updating the coefficients and the hardware resources (such as coefficient memories and number of look-up tables) in real time field-programmable gate array (FPGA) platforms is then reduced using this method. A reduction in hardware complexity by a factor of about 3 is achieved by the proposed joint method. The proposed structure is experimentally validated with a 40Gbit/s 16-QAM signal. A 7dB power penalty reduction is experimentally achieved at a bit error rate (BER) of 10^{-3} in the presence of a 10 degree phase imbalance, confirming the effectiveness of the proposed algorithm. The equalization capability remains even in the presence of group velocity dispersion along the link, which is numerically confirmed with optical fiber transmission up to 1200 km and 20 phase imbalance.

In [50], circular harmonic expansion-based carrier frequency offset estimation was investigated for optical m-QAM communication systems. The proposed method, combined with a gradient-descent algorithm, shows better performance compared to already proposed VVMFOE and 4th-power methods.

CELTIQUE Project-Team

6. New Results

6.1. Certified compilation

We thrive at improving the technology of certified compilation. Our work builds on the infrastructure provided by the CompCert compiler. We are working both at improving the guarantees provided by certified compilation and at formalising state-of-the-art optimisation techniques.

6.1.1. Safer CompCert

Participants: Sandrine Blazy, Frédéric Besson, Pierre Wilke.

The CompCert compiler is proved with respect to an abstract semantics. In previous work [52], we propose a more concrete memory model for the CompCert compiler [68]. This model gives a semantics to more programs and lift the assumption about an infinite memory. This model makes CompCert safer because more programs are captured by the soundness theorem of CompCert and because it allows to reason about memory consumption.

We are investigating the consequences this model on different compiler passes of CompCert [32]. As a sanity check, we prove formally that the existing memory model is an abstraction of our more concrete model thus validating formally the soundness of CompCert's abstract semantics of pointers. We have also port the front-end of the compiler to our new semantics and are working on the compiler back-end.

6.1.2. Verification of optimization techniques

Participants: Sandrine Blazy, Delphine Demange, Yon Fernandez de Retana, David Pichardie.

The CompCert compiler foregoes using SSA, an intermediate representation employed by many compilers that enables writing simpler, faster optimizers. In previous work [51], we have proposed a formally verified SSA-based middle-end for CompCert, addressing two problems raised by Leroy in 2009: giving an intuitive formal semantics to SSA, and leveraging its global properties to reason locally about program optimizations. Since then, we have studied in more depth the SSA-based optimization techniques with the objective to make the middle-end more realistic, in terms of the efficiency of optimizations, and to rationalize the way the correctness proofs of optimizations are conducted and structured.

First, we have studied in [34] the problem of a verified, yet efficient (i.e. as implemented in production compilers) technique for testing the dominance relation between two nodes in a control flow graph. We propose a formally verified validator of untrusted dominator trees, on top of which we implement and prove correct a fast dominance test.

Second, in [20], we implement and verify two prevailing SSA optimizations (Sparse Conditional Constant Propagation and Global Value Numbering), conducting the proofs in a unique and common sparse optimization proof framework, factoring out many of the dominance-based reasoning steps required in proofs of SSA-based optimizations. Our experimental evaluations indicate both a better precision, and a significant compilation time speedup.

Finally, we have studied (paper under review at the international conference Compiler Construction 2016) the destruction of the SSA form (i.e. at the exit point of the middle-end), a problem that has remained a difficult problem, even in a non-verified environment. We formally defined and proved the properties of the generation of Conventional SSA (CSSA) which make its destruction simple to implement and prove. We implemented and proved correct a coalescing destruction of CSSA, à la Boissinot et al., where variables can be coalesced according to a refined notion of interference. Our CSSA-based, coalescing destruction allows us to coalesce more than 99% of introduced copies, on average, and leads to encouraging results concerning spilling and reloading during post-SSA allocation.

6.2. Certified Static Analyses

6.2.1. Certified Analyses for JavaScript

Participants: Martin Bodin, Thomas Jensen, Alan Schmitt.

We have continued our work on the certification of analyses for JavaScript by developing a systematic way to build certified abstract interpreters from big-step semantics using the Coq proof assistant. We based our approach on Schmidt's abstract interpretation principles for natural semantics, and used a pretty-big-step (PBS) semantics, a semantic format proposed by Charguéraud. We proposed a systematic representation of the PBS format and implemented it in Coq. We then showed how the semantic rules can be abstracted in a methodical fashion, independently of the chosen abstract domain, to produce a set of abstract inference rules that specify an abstract interpreter. We proved the correctness of the abstract interpreter in Coq once and for all, under the assumption that abstract operations faithfully respect the concrete ones. We finally showed how to define correct-by-construction analyses: their correction amounts to proving they belong to the abstract semantics. This work has been published at CPP 2015 [19].

In addition, we have worked on hybrid typing of information flow for JavaScript, in collaboration with José Fragoso Santos and Tamara Rezk at Inria Sophia-Antipolis. Our analysis combined static and dynamic typing in order to avoid rejecting programs due to imprecise typing information. This work has been published at TGC 2015 [21].

6.2.2. Certified Analyses for safety-critical C programs

Participants: Sandrine Blazy, Vincent Laporte, David Pichardie.

We designed and proved sound, using the Coq proof assistant, a static analyzer, Verasco [26], based on abstract interpretation for most of the ISO C 1999 language (excluding recursion and dynamic allocation). Verasco establishes the absence of run-time errors in the analyzed programs. It enjoys a modular architecture that supports the extensible combination of multiple abstract domains, both relational and non-relational. Verasco integrates with the CompCert formally-verified C compiler so that not only the soundness of the analysis results is guaranteed with mathematical certitude, but also the fact that these guarantees carry over to the compiled code.

6.2.3. Certified Analyses for binary codes

Participants: Sandrine Blazy, Vincent Laporte, David Pichardie.

Static analysis of binary code is challenging for several reasons. In particular, standard static analysis techniques operate over control flow graphs, which are not available when dealing with self-modifying programs which can modify their own code at runtime. We formalized in the Coq proof assistant some key abstract interpretation techniques that automatically extract memory safety properties and control flow graphs from binary code [13], and operate over a small subset of the x86 assembly. Our analyzer is formally proved correct and has been run on several self-modifying challenges, provided by Cai et al. in their PLDI 2007 paper. This an extended version of out ITP 2014 paper.

6.3. Static analysis of functional programs using tree automata and term rewriting

Participants: Thomas Genet, 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 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. Now, we aim at applying this technique to the static analysis of programming languages whose semantics is based on terms, like functional programming languages. We already shown that static analysis of first order functional programs with a call-by-value evaluation strategy can be automated using tree automata completion [22]. This is the subject of the PhD thesis Yann Salmon has defended [11]. Now, one of the objective is to lift those results to the static analysis of higher-order functions.

6.4. Static analysis of functional specifications

Participants: Thomas Jensen, Oana Andreescu.

We have developed a static dependency analysis for a strongly typed, high-level functional specifications written in a specification formalism developed by the SME Prove & Run. In the context of interactive formal verification of complex systems, much effort is spent on proving the preservation of the system invariants. However, most operations have a localized effect on the system, which only really impacts few invariants at the same time. Identifying those invariants that are unaffected by an operation can substantially ease the proof burden for the programmer. Our dependency analysis computes a conservative approximation of the input fragments on which the operations depend. It is a flow-sensitive interprocedural analysis that handles arrays, structures and variant data types. We have validated the scalability of the analysis to complex transition systems by applying it to a functional specification of the MINIX operating system. This work was reported in [25].

6.5. Semantics

6.5.1. Energy-valued semantics

Participant: David Cachera.

We develop a *-continuous Kleene ω-algebra of real-time energy functions [36]. Together with corresponding automata, these can be used to model systems which can consume and regain energy (or other types of resources) depending on available time. Using recent results on *-continuous Kleene ω-algebras and computability of certain manipulations on real-time energy functions, it follows that reachability and Büchi acceptance in real-time energy automata can be decided in a static way which only involves manipulations of real-time energy functions. This works opens the way to static analysis techniques for energy-valued semantics of real-time systems.

DECENTRALISE Team

6. New Results

6.1. Asyncronous Messaging

There are now a variety of end-to-end encrypted messaging platforms targeted at personal correspondences. Amongst these, only Pond and Ricochet provide meaningful resistance to traffic analysis by explicitly protecting the message metadata, although several can optionally operate over Tor to protect the user's location. Ricochet's design around Tor hidden services does not permit offline operation. Pond depends upon a centralized server.

In addition, there are messengers designed for academic research, like Vuvuzela, Dissent, and DP5. These employ information theoretically secure channels like dining cryptographers networks (DC-nets) and private information retrieval schemes (PIR) because they admit extremely simply proofs of security. As DC-nets and PIR schemes scale quadratically, these messaging research projects are effectively limited to a fixed maximum number of users, so they cannot realistically provide an alternative to modern email.

Instead, we have begun exploring the prospects of using mid-latency store-and-forward mixnets for asynchronous messaging. In fact, these are the amongst oldest anonymity systems, dating back to David Chaum, but they were normally restricted to anonymous email projects. At present, we remain in the early design phase, but our design scales linearly while providing many interesting properties desired by modern messengers.

We obtain provable security by basing our system on the Sphinx mixnet packet format, which is provably secure in the universal composability framework [7]. At first blush, Sphinx appears to be an overly restrictive format, but the restrictions are worth obtaining this degree of provable security along with a mixnet's scalability. After consideration, we have devised methods for adding entropy, and optimizing the location of entropy in Sphinx packet headers, without the need to use a larger and slower elliptic curve.

In Sphinx, there is a facility for single-use reply blocks (SURBs), as in other mixnets initially designed for anonymous remailers whose forward and backward messages look alike. We can store an SURB in the packet header, which enters use when the packet passes a fixed cross-over node, thereby allowing both sender and receiver remain anonymous to one another. We can orchestrate the usage of SURBs, and an authentication scheme using tokens, to provide optimal messaging propoerties that:

- Protect the identities of senders and recipients from each other and mixnet nodes, including the mailbox servers,
- Protect the identities of recipient's mailbox servers from even their contact to prevent denial of services attack,
- All redudancy through the usage of multiple mailbox servers.

We shall employ the Axolotl ratchet for long-term forward secrecy in messages, like Pond and Signal do. We can slightly improve upon the Axolotl ratchet by judiciously introducing side key material into the ratchet state. These side keys could be symmetric keys that take a different route through the mixnet, or travel outside the mixnet, thereby allowing the ratchet state to evolve based upon multiple concurrent paths. Side keys could also employ post-quantum public key cryptography, thus providing forward-secrecy against future attackers equipped with quantum computers.

We have also found another forward-secure ratchet inspired by Axolotl that integrates well with the Sphinx packet format. We believe this allows mixnet messages to be protected by long-term ratchets and posses a modicum of protection even against attackers with quantum-computers. At best, long-term ratchets themselves are only pseudonymous, not actually anonymous, so using the integrated ratchets requires considerable care.

6.2. Efficient Privacy-Preserving Scalar Product

We have designed, implemented and evaluated two variants of new privacy-preserving scalar product protocols. The first variant is based on an original idea of Ioannidis et al. [8] and was refined by Amirbekyan et al. [6]. Our first design improves on this by supporting signed values. A second design uses discrete logarithms over Elliptic curves instead of a homomorphic cipher, resulting in a substantially more efficient computation as long as the final result is numerically small.

In both protocols, Alice learns the scalar product $\sum a_i b_i$ of her private vector $\overrightarrow{a}$ with Bob's private vector $\overrightarrow{b}$. The protocol is privacy-preserving in that Alice cannot discern details about $\overrightarrow{b}$ other than what she can learn from $\overrightarrow{a}$ and the scalar product $\sum a_i b_i$, and Bob does not learn anything.

Table 1 summarizes our experimental results.

Table 1. Preliminary performance data for the SP algorithms, wall-clock time running on a single-core of an i7.

Length	RSA-2048	ECC-2^{20}	ECC-2^{28}
25	14 s	2 s	29 s
50	21 s	2 s	29 s
100	39 s	2 s	29 s
200	77 s	3 s	30 s
400	149 s	OOR	31 s
800	304 s	OOR	33 s
800	3846 kb	OOR	70 kb

6.3. GNS support for Tor

We have worked with the Tor community to understand how best to support integration of the GNU Name System with Tor via specialized Tor exit nodes. There are two components to this work:

At present, there are somewhat fragile configuration options to Tor that should allow Tor users to locate the specialized exit nodes, although a small patch to Tor itself would improve upon these.

There are security reasons why Tor should not interact with locally configured name resolution services. OnioNS created a method to make Tor use local services for some domain name lookups, but doing so is somewhat heavy [9]. If we're creating a GNS patch to Tor anyways, then we'll likely extend it to optimize this process.

ESTASYS Team

6. New Results

6.1. Heterogeneous Systems

Participants: Axel Legay, Jean Quilbeuf.

This part concerns Tasks 1, 2 and 4 of the action. We characterize and formalize heterogeneous aspects of SoS and then we define efficient monitoring algorithms and representations for their requirements. We then combine the results with Statistical Model Checking (Task 5).

Systems of Systems (SoS) are very large scale systems with particular characteristics. SoS are not directly built from scratch by a single designer or a single team but are obtained as the composition of simpler systems. SoS have strong reliability and dependability requirements, as they aim to provide a service over a long running period. SoS may dynamically modify themselves by connecting to new systems, updating or disconnecting faulty ones, making it impossible to statically know the set of subsystems that are part of the SoS before runtime.

One of the main difficulty arising when developing SoS is the fact that subsystems may have been designed with a different goal in mind. In particular, some subsystems may have their own goal which differs from the global goal of the SoS. Furthermore, each subsystem may be developed in a particular computation model, making it difficult to find a common unifying semantics for the whole SoS. Finally, SoS may exhibit some emergent behaviors that are hardly predictable at design time.

One of the solutions to allow simulation of an SoS is to rely on a common interface for interconnecting the subsystems. The Functional Mockup Interface (FMI) standard is a natural candidate for such an interface. The different components of an SoS developed in different models of computation can be translated to Functional Mockup Units (FMU). Then a so-called master algorithm coordinates the FMUs composing the system. The execution of each FMU is either directly handled by the master algorithm or relies on an external tool for its execution.

Because the subsystems composing an SoS are of heterogeneous nature, it is difficult to find a common semantics model for the whole system. Furthermore, building such a transition system is not tractable due to the complexity of the system. Thus verification through traditional model checking is not possible for SoS. However, since the FMI/FMU framework enables simulation of such systems, the statistical model checking approach can be used.

The DANSE EU project aims to provide a complete tool chain from the modeling to the verification of SoS. At the higher level, the modeling is done in UPDM using the RHAPSODY tool. At the same level, the designer can express requirements over the model using some patterns written in GCSL. The UPDM model can then be translated into a FMI/FMU format that can be simulated by a dedicated tool, named DESYRE. Similarly, the GCSL requirements are transformed into BLTL formulas. Finally, the PLASMA statistical model checker has been integrated with the DESYRE tool chain in order to check the BLTL formulas based on the simulations provided by DESYRE.

6.1.1. Papers:

papier DANSE(en cours) Ensuring a correct behaviour of SoS has a significant social impact. Their complexity and inherent dynamicity pose a serious challenge to traditional design methodologies. We propose a methodology and a tool-chain supporting design and validation of SoSs. We integrate SMC with existing industrial practice, by addressing both methodological and technological issues. Our contribution is summarized as follows: (1) a methodology for continuous and scalable validation of SoS formal requirements; (2) a natural-language based formal specification language able to express complex SoS requirements; (3) adoption of widely used industry standards for simulation

and heterogeneous systems integration (FMI and UPDM); (4) development of a robust SMC tool-chain integrated with system design tools used in practice. We illustrate the application of our SMC tool-chain and the obtained results on an industrial case study from the DANSE project.

6.2. Statistical Model Checking

Participants: Axel Legay, Sean Sedwards, Jean Quilbeuf, Louis-Marie Traonouez, Chan Ngo, Cyrille Jegourel.

This section covers Tasks 4 and 5 of the action. It consists in developping Simulation based techniques and efficient statistical algorithms for SoS.

The use of test cases remains the default means of ensuring the correct behaviour of systems in industry, but this technique is limited by the need to hypothesise scenarios that cause interesting behaviour and the fact that a reasonable set of test cases is unlikely to cover all possible eventualities. Static analysis is more thorough and has been successful in debugging very large systems, but its ability to analyse complex dynamical properties is limited. In contrast, model checking is an exhaustive technique that verifies whether a system satisfies a dynamical temporal logic property under all possible scenarios. For nondeterministic and probabilistic systems, numerical model checking quantifies the probability that a system satisfies a property. It can also be used to quantify the expected cost or reward of sets of executions.

Numerical model checking gives precise, accurate and certain results by exhaustively exploring the state space of the model, however the exponential growth of the state space with system size (the ‘state explosion problem) typically limits its applicability to “toy” systems. Symbolic model checking using efficient data structures can make certain very large models tractable. It may also be possible to construct simpler but behaviourally equivalent models using various symmetry reduction techniques, such as partial order reduction, bisimulation and lumping. If a new system is being constructed, it may be possible to guarantee the overall behaviour by verifying the behaviour of its subcomponents and limiting the way they interact. Despite these techniques, however, the size, unpredictability and heterogeneity of real systems usually make numerical techniques infeasible. Moreover, even if a system has been specified not to misbehave, it is nevertheless necessary to check that it meets its specification.

Simulation-based approaches are becoming increasingly tractable due to the availability of high performance parallel hardware and algorithms. In particular, statistical model checking (SMC) combines the simplicity of testing with the formality of numerical model checking. The core idea of SMC is to create multiple independent execution traces of a system and count how many satisfy a property specified in temporal logic. The proportion of satisfying traces is an estimate of the probability that the system satisfies the property. By thus modelling the executions of a system as a Bernoulli random variable, the absolute error of the estimate can be bounded using, for example, a confidence interval or a Chernoff bound. It is also possible to use efficient sequential hypothesis testing, to decide with specified statistical confidence whether the probability of a property is above or below a given threshold. Since SMC requires multiple independent simulations, it may be efficiently divided on parallel computer architectures, such as grids, clusters, clouds and general purpose computing on graphics processors (GPGPU).

Knowing a result with less than 100% confidence is often sufficient in real applications, since the confidence bounds may be made arbitrarily tight. Moreover, a swiftly achieved approximation may prevent a lot of wasted time during model design. For many complex systems, SMC offers the only feasible means of quantifying performance. Historically relevant SMC tools include APMC, YMER and VESTA. Well-established numerical model checkers, such as PRISM and UPPAAL, are now also including SMC engines. Dedicated SMC tools under active development include COSMOS and our own tool PLASMA. Recognising that SMC may be applied to any discrete event trace obtained by stochastic simulation, we have devised PLASMA-lab, a modular library of SMC algorithms that may be used to construct domain-specific SMC tools. PLASMA-lab has become the main vehicle of our ongoing development of SMC algorithms.

Statistical model checking (SMC) addresses the state explosion problem of numerical model checking by estimating quantitative properties using simulation. To advance the state of the art of SMC we address the ongoing challenges of rare events and nondeterminsm. We also make novel use of SMC by applying it to motion planning in the context of assisted living. Rare events are often of critical importance and are challenging to SMC because they appear infrequently in simulations. Nondeterministic models are useful to model unspecified interactions, but simulation requires that nondeterminism is resolved.

We also applied SMC in the context of Systems of Systems (SoS). In the frame of the DANSE project, Plasma-Lab was used to verify SoS, and completely integrated with the DANSE tool-chain. We are currently working on verification of dynamic SoS, where systems can appear and disappear during execution. This work is done in collaboration with the ArchWare team from IRISA. We will interface Plasma-Lab with a simulator for the Pi-ADL language that enables simulation of dynamic systems.

Our group is devising cutting edge techniques for SMC. In particular, we are developing new algorithms for non-deterministic systems as well as for dynamic systems. Rare event systems are also addressed. Finally, we also devote a large amount of time to applying our technology to realistic case studies described in high-level languages such as Simulink or System C, or even a robot moving an eldery person in a commercial center.

6.2.1. Papers:

[2] (J) People with impaired physical and mental ability often find it challenging to negotiate crowded or unfamiliar environments, leading to a vicious cycle of deteriorating mobility and sociability. To address this issue we present a novel motion planning algorithm that is able to intelligently deal with crowded areas, permanent or temporary anomalies in the environment (e.g., road blocks, wet floors) as well as hard and soft constraints (e.g., “keep a toilet within reach of 10 meters during the journey”, “always avoid stairs”). Constraints can be assigned a priority tailored on the user’s needs. The planner has been validated by means of simulations and experiments with elderly people within the context of the DALi FP7 EU project.

[3] (J) Markov decision processes (MDP) are useful to model optimisation problems in concurrent systems. To verify MDPs with efficient Monte Carlo techniques requires that their nondeterminism be resolved by a scheduler. Recent work has introduced the elements of lightweight techniques to sample directly from scheduler space, but finding optimal schedulers by simple sampling may be inefficient. Here we describe “smart” sampling algorithms that can make substantial improvements in performance.

[21] (C) Rare properties remain a challenge for statistical model checking (SMC) due to the quadratic scaling of variance with rarity. We address this with a variance reduction framework based on lightweight importance splitting observers. These expose the model-property automaton to allow the construction of score functions for high performance algorithms. The confidence intervals defined for importance splitting make it appealing for SMC, but optimising its performance in the standard way makes distribution inefficient. We show how it is possible to achieve equivalently good results in less time by distributing simpler algorithms. We first explore the challenges posed by importance splitting and present an algorithm optimised for distribution. We then define a specific bounded time logic that is compiled into memory-efficient observers to monitor executions. Finally, we demonstrate our framework on a number of challenging case studies.

[23] (C) Exhaustive verification can quantify critical behaviour arising from concurrency in nondeterministic models. Rare events typically entail no additional challenge, but complex systems are generally untractable. Recent work on Markov decision processes allows the extremal probabilities of a property to be estimated using Monte Carlo techniques, offering the potential to handle much larger models. Here we present algorithms to estimate extremal rewards and consider the challenges posed by rarity. We find that rewards require a different interpretation of confidence and that reachability rewards require the introduction of an auxiliary hypothesis test. We show how importance sampling can significantly improve estimation when probabilities are low, but find it is not a panacea for rare schedulers.

[36] (J; accepted) We propose a new SMC technique based on CUSUM, an algorithm originally used in signal processing, that detects probability change at runtime on a single execution of a system. The principle is to monitor the execution at regular time intervals, and to perform Monte Carlo checks over the samples of the execution. The results of these checks are used to compute the CUSUM ratio, whose variation allows to detect a change of the probability measure of the system. We demonstrate the algorithm to detect failures in a Simulink model of a temperature controller. Computing the exact time at which failures may happen is then useful to schedule maintenance operations.

[42] (W) Many embedded and real-time systems have a inherent probabilistic behaviour (sensors data, unreliable hardware,...). In that context, it is crucial to evaluate system properties such as "the probability that a particular hardware fails". Such properties can be evaluated by using probabilistic model checking. However, this technique fails on models representing realistic embedded and real-time systems because of the state space explosion. To overcome this problem, we propose a verification framework based on *Statistical Model Checking*. Our framework is able to evaluate probabilistic and temporal properties on large systems modelled in SystemC, a standard system-level modelling language. It is fully implemented as an extension of the Plasma-lab statistical model checker. We illustrate our approach on a multi-lift system case study.

[27] (W) Stochastic Petri nets are commonly used for modeling distributed systems in order to study their performance and dependability. This report proposes a realization of stochastic Petri nets in SystemC for modeling large embedded control systems. Then statistical model checking is used to analyze the dependability of the constructed model. Our verification framework allows users to express a wide range of useful properties to be verified which is illustrated through a case study.

[25] (C: accepted) Transaction-level modeling with SystemC has been very successful in describing the behavior of embedded systems by providing high-level executable models, in which many of them have an inherent probabilistic behavior, i.e., random data, unreliable components. It is crucial to evaluate the quantitative and qualitative analysis of the probability of the system properties. Such analysis can be conducted by constructing a formal model of the system and using probabilistic model checking. However, this method is infeasible for large and complex systems due to the state space explosion. In this work, we demonstrate the successful use of *Statistical Model Checking* to carry out such analysis directly from large SystemC models and allows designers to express a wide range of useful properties. This work is going to presented at 17th IEEE High Assurance Systems Engineering Symposium in January, 2016.

6.3. Formal Models for Variability

Participants: Axel Legay, Rudolf Fahrenberg, Jin Hyun Kim.

> This part of the report is more concerned with task 2. It studies variability aspects in the broad scope. As in the first year, we have decided to use the concept of product lines as a general framework to reason on the problematic.

The behaviour of a software system is often described in terms of its features, where each *feature* is a unit of functionality that adds value to the system. *Feature-oriented software development (FOSD)* is a software-development strategy that is based on feature decomposition and modularity. Features can be separate modules that are developed in isolation, allowing for parallel, incremental, or multi-vendor development of features. Feature orientation is particularly important in *software product lines*, where a family of related products is managed and evolved in terms of its features: a product line comprises a collection of mandatory and optional features, and individual products are derived by selecting among and integrating features from this feature set. A product line can be expressed as a single model, in which feature-specific behaviour is conditional on the presence of the feature in a product.

The downside of FOSD is that, although features are conceptualized, developed, managed, and evolved as separate concerns, they are not truly separate. They can interfere with each other, for example by trying to control the same variables, by issuing events that trigger other features, or by imposing conditions that suppress other features. Most of the early work on feature interactions focused on interactions that manifest themselves as logical inconsistencies, such as conflicting actions, nondeterminism, deadlock, invariant violation, or unsatisfiability. More recently, a more general definition of feature interaction has been introduced, in terms of a feature that is developed and verified to be correct in isolation but is found to behave differently when combined with other features, and it was shown how such *behaviour interactions* could be detected as a violation of bisimulation.

Another problem is that FTS models are monolithic models of full product lines. There is no means of modelling individual features and composing them into products or product-line models, or of specifying feature increments to an existing product-line model. As such, FTSs cannot be the mathematical basis for modelling technologies that support feature decomposition, composition, or incremental evolution of a product line.

6.3.1. Papers:

[11] (C) Featured Transition Systems (FTSs) is a popular representation for software product lines: an entire product line is compactly represented as a single transition-machine model, in which feature-specific behaviour is guarded by feature expressions that are satisfied (or not) by the presence or absence of individual features. In previous work, FTS models were monolithic in the sense that the modeller had to construct the full FTS model of the product line in its entirety. To allow for modularity of FTS models, we propose here a language for extending an existing FTS model with new features. We demonstrate the language using a running example and present results about the language's expressivity, commutativity of feature extensions, feature interactions, and resolution of such interactions.

[12] (C) We suggest a method for measuring the degree to which features interact in feature-oriented software development. To this end, we extend the notion of simulation between transition systems to a similarity measure and lift it to compute a behaviour interaction score in featured transition systems. We then develop an algorithm which can compute the degree of feature interactions in a featured transition system in an efficient manner.

6.4. Privacy and Security

Participants: Axel Legay, Fabrizio Biondi, Jean Quilbeuf, Thomas Given-Wilson, Sébastien Josse.

6.4.1. Information-Theoretical Quantification of Security Properties

This part of the work was not foreseen at the beginning of the action. It concerns security aspects, and more precisely quantifying privacy of data. This aspect is in fact central for SoS and all our algorithms developed for Tasks 4 and 5 should be adapted to solve a series of problems linked to privacy in interconnected object and dynamical environment. For now, we only studied the foundations.

Information theory provides a powerful quantitative approach to measuring security and privacy properties of systems. By measuring the *information leakage* of a system security properties can be quantified, validated, or falsified. When security concerns are non-binary, information theoretic measures can quantify exactly how much information is leaked. The knowledge of such informations is strategic in the developments of component-based systems.

The quantitative information-theoretical approach to security models the correlation between the secret information of the system and the output that the system produces. Such output can be observed by the attacker, and the attacker tries to infer the value of the secret by combining this information with its knowledge of the system.

Armed with the produced output and the source code of the system, the attacker tries to infer the value of the secret. The quantitative analysis we implement computes with arbitrary precision the number of bits of the secret that the attacker will expectedly infer. This expected number of bits is the information leakage of the system.

The quantitative approach generalizes the qualitative approach and thus provides superior analysis. In particular, a system respects non-interference if and only if its leakage is equal to zero. In practice very few systems respect non-interference, and for those who dont it is imperative to be able to distinguish between the ones leaking a very small amount of bits and the ones leaking a significant amount of bits, since only the latter are considered to pose a security vulnerability to the system.

Since black box security analyzes are immediately invalidated whenever an attacker gains information about the source code of the system, we assume that the attacker has a white box view of the system, meaning that it has access to the systems source code. This approach is also consistent with the fact that many security protocol implementations are in fact open source.

The scope of modern software projects is too large to be analyzed manually. For this reason we provide tools that can support the analyst and locate security vulnerabilities in large codebases and projects. We work with a variety of tools, including commercial software analysis tools being adapted with our techniques, and tools such as QUAIL developed here by our team.

We applied the leakage analysis provided by QUAIL to several case studies. Our case studies (voting protocol and smart grid coordination) have in common that a publicly disclosed information is computed from the secret of every participant in the model. In the voting example, the vote of a given voter is secret, but the number of votes for each candidates is public. Similarly, in the smart grid example, the consumption of one of the houses is secret, but the consumption of a whole quarter can be deduced. Qualitative analyses are either too restrictive or too permissive on these types of systems. For instance, non-interference will reject them as the public information depends on the secret. Declassification approaches will accept them, even if the number of voters or consumers is 2, in which case the secret can be deduced.

The development of better tools for quantitative security builds upon both theoretical developments in information theory, and development of the tools themselves. These often progress in parallel with each supporting the findings of the other, and increasing the demands and understanding upon each other.

6.4.1.1. Papers:

[34] (C; submitted) Systems dealing with confidential data may leak some information by their observable outputs. Quantitative information flow analysis provides a method for quantifying the amount of such information leakage. To avoid the high computational cost of exhaustive search, statistical analysis has been studied to estimate information leakage by analyzing only a small but representative subset of the system's behavior. In this paper we propose a new compositional statistical analysis method for quantitative information flow that combines multiple statistical analyses with static trace analysis. We use partial knowledge of the system's source code or specification, therefore improving both quality and cost of the analysis. The new method can optimize the use of weighted statistical analysis by performing it on components of the system and appropriately adapting their weights. We show this approach combined with the precision of trace analysis produces better estimates and narrower confidence intervals than the state of the art.

[38] (J) The quantification of information leakage provides a quantitative evaluation of the security of a system. We propose the usage of Markovian processes to model deterministic and probabilistic systems. By using a methodology generalizing the lattice of information approach we model refined attackers capable to observe the internal behavior of the system, and quantify the information leakage of such systems. We also use our method to obtain an algorithm for the computation of channel capacity from our Markovian models. Finally, we show how to use the method to analyze timed and non-timed attacks on the Onion Routing protocol.

[40] (C) Quantitative security analysis evaluates and compares how effectively a system protects its secret data. We introduce QUAIL, the first tool able to perform an arbitrary-precision quantitative

analysis of the security of a system depending on private information. QUAIL builds a Markov Chain model of the system's behavior as observed by an attacker, and computes the correlation between the system's observable output and the behavior depending on the private information, obtaining the expected amount of bits of the secret that the attacker will infer by observing the system. QUAIL is able to evaluate the safety of randomized protocols depending on secret data, allowing to verify a security protocol's effectiveness. We experiment with a few examples and show that QUAIL's security analysis is more accurate and revealing than results of other tools.

[41] (C) Quantitative security techniques have been proven effective to measure the security of systems against various types of attackers. However, such techniques are based on computing exponentially large channel matrices or Markov chains, making them impractical for large programs. We propose a different approach based on abstract trace analysis. By analyzing directly sets of execution traces of the program and computing security measures on the results, we are able to scale down the exponential cost of the problem. Also, we are able to appy statistical simulation techniques, allowing us to obtain significant results even without exploring the full space of traces. We have implemented the resulting algorithms in the QUAIL tool. We compare their effectiveness against the state of the art LeakWatch tool on two case studies: privacy of user consumption in smart grid systems and anonymity of voters in different voting schemes.

[37] (C) In an election, it is imperative that the vote of the single voters remain anonymous and undisclosed. Alas, modern anonymity approaches acknowledge that there is an unavoidable leak of anonymity just by publishing data related to the secret, like the election's result. Information theory is applied to quantify this leak and ascertain that it remains below an acceptable threshold. We apply modern quantitative anonymity analysis techniques via the state-of-the-art QUAIL tool to the voting scenario. We consider different voting typologies and establish which are more effective in protecting the voter's privacy. We further demonstrate the effectiveness of the protocols in protecting the privacy of the single voters, deriving an important desirable property of protocols depending on composite secrets.

[39] (C) In recent years, quantitative security techniques have been providing effective measures of the security of a system against an attacker. Such techniques usually assume that the system produces a finite amount of observations based on a finite amount of secret bits and terminates, and the attack is based on these observations. By modeling systems with Markov chains, we are able to measure the effectiveness of attacks on non-terminating systems. Such systems do not necessarily produce a finite amount of output and are not necessarily based on a finite amount of secret bits. We provide characterizations and algorithms to define meaningful measures of security for non-terminating systems, and to compute them when possible. We also study the bounded versions of the problems, and show examples of non-terminating programs and how their effectiveness in protecting their secret can be measured.

6.4.2. Equivocation-based Security Measures for Shared-Key Cryptosystems

Ensuring privacy and security of communication is a fundamental concern of cyber-physical systems and handled by encryption. Information-theoretic reasoning allows the modelling of security properties via unconditional security. That is, information-theoretic approaches formalise security properties that do not rely upon unproven computational hardness results, and are not vulnerable to advances in computing hardware, software or theory. For example, such unconditional security guarantees are not weakened by quantum computers, mem-computers, or new mathematical discoveries.

Traditionally the strongest measure of the security of a system is *perfect secrecy* as proposed by Shannon. However, this relies upon having a large key that is used only once. In practice a measure of the security of cryptosystems that does not meet this requirement is more useful. To this end we presented *max-equivocation*, a measure of the maximum achievable security given the keys available. Indeed max-equivocation not only formalizes the best possible security, but also generalizes perfect secrecy.

Max-equivocation holds even when inputs to the systems (i.e. keys and messages) are not uniform. This corresponds to many real world scenarios, and indeed we have shown that existing approaches are non-optimal as they do not consider these perturbations in the inputs. We provide necessary and sufficient conditions for achieving max-equivocation, formalizing exactly when it can be achieved in practice.

We further generalize to consider scenarios where message spaces are not complete, i.e. there are messages that are invalid and could never be produced. This allows reasoning over (and contrasting with) many prior approaches as well as formalizing their strengths and weaknesses under max-equivocation.

The most common attack against such cryptosystems is to consider when the attacker sees a single (encrypted) message and tries to guess the content. This can be measured by the *vulnerability* of the system, i.e. the probability that the attacker will guess correctly the message. We formalize a *relative vulnerability* for when the attacker makes this guess under incorrect assumptions about the messages. We formalize that the attacker can never improve their chances at guessing the message with incorrect assumptions.

Now we consider what information the attacker can gain by observing the cryptosystem. We show that the encryption function alone reveals information about the possible message distributions to the attacker. In the worse case scenario an encryption function may admit only a single message distribution. Thus the construction of the encryption function should consider this and (when possible) admit many solutions.

Further we consider what the attacker can learn by observing the communication of a cryptosystem. We show that the attacker can learn the probability distribution over the ciphertexts (encrypted messages), and combined with the information from the encryption function converge upon a distribution for the messages. Again if the encryption function admits one solution then the attacker learns the exact message distribution. We show that even when a single solution will not be found, the attacker still converges upon a message distribution that can only improve their attacks.

In addition to formalizing how these attacks work, and thus how to protect against them when constructing cryptosystems, we also consider not sharing the encryption function as a mechanism to avoid the attacker exploiting it. We formalize how to still communicate effectively in this scenario, and the advantages and disadvantages of this approach.

We present several algorithms to demonstrate the practicality of the techniques. The algorithms to achieve max-equivocation consider the message distribution and compute an encryption function that achieves close to max-equivocation. Since these algorithms are tailored for the message distributions, they out perform generic algorithms. We also present algorithms that are able to perform well without revealing the entire encryption function, and thus revealing less information to the attacker and hindering cryptoanalysis.

Thus we show that unconditional security is not only more resistant to technology changes, but also can be formalised for many scenarios, and is achievable in practice.

6.4.2.1. Papers:

[29] (C, submitted) Recent work has presented max-equivocation as a measure of the resistance of a cryptosystem to attacks when the attacker is aware of the encoder function and message distribution. Here we consider the vulnerability of a cryptosystem in the one-try attack scenario when the attacker has incomplete information about the encoder function and message distribution. We show that encoder functions alone yield information to the attacker, and combined with inferable information about the ciphertexts, information about the message distribution can be discovered. We show that the whole encoder function need not be fixed or shared a priori for an effective cryptosystem, and this can be exploited to increase the equivocation over an a priori shared encoder. Finally we present two algorithms that operate in these scenarios and achieve good equivocation results, ExPad that demonstrates the key concepts, and ShortPad that has less overhead than ExPad.

[13], [28] (C; J, submitted) Preserving the privacy of private communication is a fundamental concern of computing addressed by encryption. Information-theoretic reasoning models unconditional security where the strength of the results is not moderated by computational hardness or unproven results. Perfect secrecy is often considered the ideal result for a cryptosystem, where knowledge of the ciphertext reveals no information about the message or key, however often this is impossible to

achieve in practice. An alternative measure is the equivocation, intuitively the average number of message/key pairs that could have produced a given ciphertext. We show a theoretical bound on equivocation called max-equivocation and show that this generalizes perfect secrecy when achievable, and provides an alternative measure when perfect secrecy is not. We derive bounds for max-equivocation, and show that max-equivocation is achieved when the entropy of the ciphertext is minimized. We consider encryption functions under this new perspective, and show that in general the theoretical best is unachievable, and that some popular approaches such as Latin squares or Quasigroups are also not optimal. We present some algorithms for generating encryption functions that are practical and achieve 90 - 95% of the theoretical best, improving with larger message spaces.

6.4.3. Malware Classification via Deobfuscation and Behavioral Fingerprinting

A fundamental problem to guarantee the security of systems is to be able to discriminate between legitimate processes and processes with malicious behavior. Malicious software, or malware, has to be identified and prevented from executing on the system, and its actions reverted by a disinfection process. To be able to recognize and disinfect malware it is necessary to be able to extract a behavioral fingerprint or signature from a binary file, and to construct a database of such signatures for comparison. The signatures in the database have to be classified accoring to the malware's family and category, allowing the correct disinfection method to be deployed.

Automatic extraction of behavioral signatures in the form of temporal logical graphs or control flow graphs is a recent but very effective technique, and malware developers have already adapted malware compilation chains to include techniques to hinder reverse engineering and thus prevent the extraction of such signatures. These obfuscation techniques include the addition of obfuscated conditional statements leading to dead code, control flow flattening based on complex function like cryptographic hash functions, and source code virtualization on an embedded interpreter.

Consequently, deobfuscation has to be developed along with fingerprinting techniques to be able to effectively extract malware signatures. We are pushing the state of the art in both subjects, advancing generalized and targeted deobfuscation and deploying them on an innovative virtualization and malware fingerprinting tool.

Mixed Boolean Arithmetic (MBA) obfuscation is an obfuscation technique developed by Cloakware Inc. and deployed in obfuscating compilation chains for both legitimate code and malware. We have deployed state-of-the-art SMT solvers to evaluate their effectiveness against MBA-obfuscated conditionals and ascertained their limited effectiveness. So we have developed an algebraic simplification technique targeting the algebraic structure of MBA obfuscation, and proved such technique to be extremely effective, being able to deobfuscate statements in orders of magnitude less time than the time required to obfuscate them in the first place.

While the algebraic simplification technique is very effective against MBA obfuscation, it is completely tailored to MBA obfuscation. We instead explore a completely general method based on dynamic program synthesis. Synthesis algorithms, like the ones based on Reed-Muller expansion techniques, interrogate the target (in this case the obfuscated conditional) multiple times considering it as a black-box oracle, and synthesize the function expressed by the target frm the answers to such interrogation. We determined that synthesis is significantly more efficient than SMT solving in synthesizing the obfuscated function in a very compact form, and thus very promising as a generalized deobfuscation method.

While more targeted deobfuscation techniques are required to coutneract control flow flattening and virtualization, the deobfuscation of conditional statements is an important step for malware fingerprinting. We plan to use our tool to classify a large database of malware, producing an extensive database of malware signatures representing multiple versions and families of malicious code. Malware mining and evolution techniques can be deployed on such database to construct different signatures for unknown variants of similar malware, thus improving the effectiveness of the detection process.

6.4.3.1. Papers:

[30] (C, submitted) The obfuscation of conditional statements is a simple and efficient way to disturb the identification of the control flow graph of a program. Mixed Boolean arithmetics (MBA) techniques provide concrete ways to achieve this obfuscation of conditional statements. In this work, we

study the effectiveness of automated deobfuscation of MBA obfuscation, using algebraic, SMT-based and synthesis-based techniques. We experimentally ascertain the practical feasibility of MBA obfuscation. We study using SMT-based approaches with different state-of-the-art SMT solvers to counteract MBA obfuscation, and we show how the deobfuscation complexity can be greatly reduced by algebraic simplification. We also consider synthesis-based deobfuscation and find it to be more effective than SMT-based deobfuscation. We discuss complexity and limits of all methods, and conclude that MBA obfuscation is not effective enough to be considered a reliable method for control flow or white-box obfuscation.

6.5. Energy-Centric Systems

Participants: Axel Legay, Uli Fahrenberg.

This part is concerned with Tasks 1 and 2. Mostly, we focus on quantifying properties of interconnected objects such as Cyber Physical Systems (CPS) (SoS and CPS share a lot of commonalities).

Energy and *resource management* problems are important in areas such as embedded systems or autonomous systems. They are concerned with the question whether a given system admits infinite schedules during which (1) certain tasks can be repeatedly accomplished and (2) the system never runs out of energy (or other specified resources). Formal modeling and analysis of such problems has attracted some attention in recent years.

6.5.1. Papers:

[18] (C; accepted) We define and study basic properties of *-continuous Kleene ω-algebras that involve a *-continuous Kleene algebra with a *-continuous action on a semimodule and an infinite product operation that is also *-continuous. We show that *-continuous Kleene ω-algebras give rise to iteration semiring-semimodule pairs, and that for Büchi automata over *-continuous Kleene ω-algebras, one can compute the associated infinitary power series.

[17] (C; accepted) Energy problems are important in the formal analysis of embedded or autonomous systems. Using recent results on *-continuous Kleene ω-algebras, we show here that energy problems can be solved by algebraic manipulations on the transition matrix of energy automata. To this end, we prove general results about certain classes of finitely additive functions on complete lattices which should be of a more general interest.

[15] (C; accepted) We develop a *-continuous Kleene ω-algebra of real-time energy functions. Together with corresponding automata, these can be used to model systems which can consume and regain energy (or other types of resources) depending on available time. Using recent results on *-continuous Kleene ω-algebras and computability of certain manipulations on real-time energy functions, it follows that reachability and Büchi acceptance in real-time energy automata can be decided in a static way which only involves manipulations of real-time energy functions.

6.6. Languages for composition

Participants: Axel Legay, Thomas Given-Wilson.

This part is concerned with Task 1, especially to describe the composition of complex systems, and to study expressivity of existing formalisms.

Contemporary cyber-physical systems are inherently constructed out of a variety of agents with communication and interaction forming a key role in the behaviour of the system as a whole. Traditional approaches to reasoning over a single computation or treating the system as a single agent prove unsatisfactory for understanding the capabilities, strengths, and weaknesses of such systems.

Since communication is a fundamental to such systems it is necessary to understand the role the communication primitives themselves play. There are many approaches to communication primitives, often chosen for their ability to easily represent desired behaviour. However, the formal properties of many implementations or chosen models have not been presented.

An alternative to formalising each possible model individually is to abstract away and reason over families of models based on their communication primitives. This allows keys results to be achieved in one model, and then generalised to the entire family, or transferred to other families based upon formal relations between these families. Thus making it possible for results to be easily applied to many models or systems without repeating significant effort.

6.6.1. Papers:

[20] (C), [32] (J; submitted) The expressiveness of communication primitives has been explored in a common framework based on the π-calculus by considering four features: synchronism (asynchronous vs synchronous), arity (monadic vs polyadic data), communication medium (shared dataspaces vs channel-based), and pattern-matching (binding to a name vs testing name equality vs intensionality). Here another dimension coordination is considered that accounts for the number of processes required for an interaction to occur. Coordination generalises binary languages such as π-calculus to joining languages that combine inputs such as the Join Calculus and general rendezvous calculus. By means of possibility/impossibility of encodings, this paper shows coordination is unrelated to the other features. That is, joining languages are more expressive than binary languages, and no combination of the other features can encode a joining language into a binary language. Further, joining is not able to encode any of the other features unless they could be encoded otherwise.

[33] (C; submitted) The expressiveness of communication primitives has been explored in a common framework by considering four features: synchronism, arity, communication medium, and pattern-matching. These all assume asymmetric communication between input and output primitives, however some calculi consider more symmetric approaches to communication such as fusion calculus and Concurrent Pattern Calculus. Symmetry can be considered either as allowing a mixture of input and output in an action or co-action, or as the unification of actions. By means of possibility/impossibility of encodings, this paper shows that: the action and co-action approach is related to or more expressive than many previously considered languages; and the unification approach is more expressive than some, but mostly unrelated to other languages.

HYCOMES Team

6. New Results

6.1. Embedded Systems Design

6.1.1. Loosely Time-Triggered Architectures: Improvements and Comparisons

Participant: Albert Benveniste.

Loosely Time-Triggered Architectures (LTTAs) are a proposal for constructing distributed embedded control systems. They build on the quasi-periodic architecture, where computing units execute 'almost periodically', by adding a thin layer of middleware that facilitates the implementation of synchronous applications. In [7], we have shown how the deployment of a synchronous application on a quasi-periodic architecture can be modeled using a synchronous formalism. Then we have detailed two protocols, Back-Pressure LTTA, reminiscent of elastic circuits, and Time-Based LTTA, based on waiting. Compared to previous work, we presented controller models that can be compiled for execution and a simplified version of the Time-Based protocol. We also compared the LTTA approach with architectures based on clock synchronization.

6.2. Hybrid Systems Modeling

Participants: Ayman Aljarbouh, Albert Benveniste, Benoît Caillaud, Khalil Ghorbal.

6.2.1. Robust Simulation for Hybrid Systems: Chattering Path Avoidance

The sliding mode approach is recognized as an efficient tool for treating the chattering behavior in hybrid systems. However, the amplitude of chattering, by its nature, is proportional to magnitude of discontinuous control. A possible scenario is that the solution trajectories may successively enter and exit as well as slide on switching mani-folds of different dimensions. Naturally, this arises in dynamical systems and control applications whenever there are multiple discontinuous control variables. The main contribution of [9] is to provide a robust computational framework for the most general way to extend a flow map on the intersection of p intersected $(n-1)$-dimensional switching manifolds in at least p dimensions. We explored a new formulation to which we can define unique solutions for such particular behavior in hybrid systems and investigate its efficient computation/simulation. An extended version of this work has been presented at the Baltic Young Scientists Conference [8].

6.2.2. A Hierarchy of Proof Rules for Checking Positive Invariance of Algebraic and Semi-Algebraic Sets

In [6], we studied sound proof rules for checking positive invariance of algebraic and semi-algebraic sets, that is, sets satisfying polynomial equalities and those satisfying finite boolean combinations of polynomial equalities and inequalities, under the flow of polynomial ordinary differential equations. Problems of this nature arise in formal verification of continuous and hybrid dynamical systems, where there is an increasing need for methods to expedite formal proofs. We study the trade-off between proof rule generality and practical performance and evaluate our theoretical observations on a set of benchmarks. The relationship between increased deductive power and running time performance of the proof rules is far from obvious; we discuss and illustrate certain classes of problems where this relationship is interesting.

6.2.3. A Formally Verified Hybrid System for Safe Advisories in the Next-Generation Airborne Collision Avoidance System

The Next-Generation Airborne Collision Avoidance System (ACAS X) is intended to be installed on all large aircraft to give advice to pilots and prevent mid-air collisions with other aircraft. It is currently being developed by the Federal Aviation Administration (FAA). In [16] we determined the geometric configurations under which the advice given by ACAS X is safe under a precise set of assumptions and formally verify these configurations using hybrid systems theorem proving techniques. We considered subsequent advisories and showed how to adapt our formal verification to take them into account. We examined the current version of the real ACAS X system and discussed some cases where our safety theorem conflicts with the actual advisory given by that version, demonstrating how formal, hybrid systems proving approaches are helping to ensure the safety of ACAS X. Our approach is general and could also be used to identify unsafe advice issued by other collision avoidance systems or confirm their safety.

6.2.4. Domain Globalization: Using Languages to Support Technical and Social Coordination

When a project is realized in a globalized environment, multiple stakeholders from different organizations work on the same system. Depending on the stakeholders and their organizations, various (possibly overlapping) concerns are raised in the development of the system. In this context a Domain Specific Language (DSL) supports the work of a group of stakeholders who are responsible for addressing a specific set of concerns. We contributed to a book chapter [11], identifying the open challenges arising from the coordination of globalized domain-specific languages. We identified two types of coordination: technical coordination and social coordination. After presenting an overview of the current state of the art, we discussed first the open challenges arising from the composition of multiple DSLs, and then the open challenges associated to the collaboration in a globalized environment.

6.3. Contracts for Systems Design

Participants: Albert Benveniste, Benoît Caillaud.

6.3.1. Contracts for Systems Design: Theory, Methodology and Application Cases

Aircrafts, trains, cars, plants, distributed telecommunication military or health care systems, and more, involve systems design as a critical step. Complexity has caused system design times and costs to go severely over budget so as to threaten the health of entire industrial sectors. Heuristic methods and standard practices do not seem to scale with complexity so that novel design methods and tools based on a strong theoretical foundation are sorely needed. Model-based design as well as other methodologies such as layered and compositional design have been used recently but a unified intellectual framework with a complete design flow supported by formal tools is still lacking. Recently an “orthogonal” approach has been proposed that can be applied to all methodologies introduced thus far to provide a rigorous scaffolding for verification, analysis and abstraction/refinement: contract-based design. Several results have been obtained in this domain but a unified treatment of the topic that can help in putting contract-based design in perspective is missing. We have published two research reports [13], [12], that intend to provide such treatment where contracts are precisely defined and characterized so that they can be used in design methodologies such as the ones mentioned above with no ambiguity. In addition, the first report [13] provides an important link between interface and contract theories to show similarities and correspondences. This report is complemented by a companion report [12] where contract based design is illustrated through use cases.

6.3.2. Contracts for Schedulability Analysis

In [10] we proposed a framework of Assume / Guarantee contracts for schedulability analysis. Unlike previous work addressing compositional scheduling analysis, our objective is to provide support for the OEM / supplier subcontracting relation. The adaptation of Assume / Guarantee contracts to schedulability analysis requires some care, due to the handling of conflicts caused by shared resources. We illustrate our framework in the context of Autosar methodology now popular in the automotive industry sector.

SUMO Project-Team

7. New Results

7.1. Model expressivity and quantitative verification

7.1.1. Diagnosability of stochastic systems

Participants: Nathalie Bertrand, Engel Lefaucheux.

Diagnosis of partially observable stochastic systems prone to faults was introduced in the late nineties. Diagnosability, i.e. the existence of a diagnoser, may be specified in different ways: (1) exact diagnosability (called A-diagnosability) requires that almost surely a fault is detected and that no fault is erroneously claimed while (2) approximate diagnosability (called ϵ-diagnosability) allows a small probability of error when claiming a fault and (3) accurate approximate diagnosability (called AA-diagnosability) requires that this error threshold may be chosen arbitrarily small. In [32] we mainly focus on approximate diagnoses. We first refine the almost sure requirement about finite delay introducing a uniform version and showing that while it does not discriminate between the two versions of exact diagnosability this is no more the case in approximate diagnosis. Then we establish a complete picture for the decidability status of the diagnosability problems: (uniform) ϵ-diagnosability and uniform AA-diagnosability are undecidable while AA-diagnosability is decidable in PTIME, answering a longstanding open question.

7.1.2. Probabilistic model checking

Participants: Blaise Genest, Ocan Sankur.

In [16], 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 symbols, we proved that the language of trajectories of distributions (one for each initial distribution) is not regular in general, even with 3 states. We then proposed a parametrized algorithm which approximates what happens to infinity, such that each symbolic block in the approximate language is at most ϵ away from the concrete distribution. We proved in [26] that if the eigenvalues of the Markov chain are distinct positive real numbers, then the trajectory is effectively regular. This is however not the case anymore if the eigenvalues can be distinct roots of real numbers.

Markov decision processes (MDPs) with multi-dimensional weights are useful to analyze systems with multiple objectives that may be conflicting and require the analysis of trade-offs. In [40], we study the complexity of percentile queries in such MDPs and give algorithms to synthesize strategies that enforce such constraints. Given a multi-dimensional weighted MDP and a quantitative payoff function f, thresholds v_i (one per dimension), and probability thresholds α_i, we show how to compute a single strategy to enforce that for all dimensions i, the probability of outcomes ρ satisfying $f_i(\rho) \geq v_i$ is at least α_i. We consider classical quantitative payoffs from the literature (sup, inf, lim sup, lim inf, mean-payoff, truncated sum, discounted sum). Our work extends to the quantitative case the multi-objective model checking problem studied by Etessami et al. [48] in unweighted MDPs.

In the invited contribution [25], we revisit the stochastic shortest path problem, and show how recent results allow one to improve over the classical solutions: we present algorithms to synthesize strategies with multiple guarantees on the distribution of the length of paths reaching a given target, rather than simply minimizing its expected value. The concepts and algorithms that we propose here are applications of more general results that have been obtained recently for Markov decision processes and that are described in a series of recent papers, including [40].

7.1.3. Stochastic modeling of biological systems

Participants: Blaise Genest, Éric Fabre, Sucheendra Palaniappan, Matthieu Pichené.

In [47], we model a population of Hela cells with non deterministic behavior, subject to the drug TRAIL. TRAIL kills a large fraction of cancerous Hela cells by triggering the apoptosis pathway. Modelling this survival is important to perform *in silico* computations helping designing treatments killing the largest fraction of cancerous cells. We model this system using the stochastic class of Dynamic Bayesian Networks. We maintain large conditional probability tables which are represented by sparse datastructure, and perform simulations by looking ahead one time step and factoring this information to avoid empty probability entries. This considerably improves the simulation based inference of DBNs, getting a 100 times improvement in its efficiency.

7.1.4. Robustness of timed models

Participants: Ocan Sankur, Loïc Hélouët.

Robustness of timed systems aims at studying whether infinitesimal perturbations in clock values can result in new discrete behaviors. A model is robust if the set of discrete behaviors is preserved under arbitrarily small (but positive) perturbations. This year we tackled this problem both for Timed Automata and time Petri Nets.

Timed automata are an extension of finite automata with clock variables that can conveniently model real-time systems. In [42], we study the robustness analysis problem for timed automata under guard imprecisions which consists in computing a timing imprecision bound under which a given specification holds. This is a particular kind of parameter synthesis problems specialized for analyzing robustness. We give a symbolic semi-algorithm for the problem based on a parametric data structure, and evaluate its performance in comparison with a recently published one, and with a binary search on the imprecision bound. We show that a safe bound on imprecision can be computed efficiently, and a performance close to that of exact model checking can be obtained thanks to the use of the parametric data structure and cycle acceleration techniques.

Another related problem is that of robust controller synthesis for timed automata where the goal is to choose actions and their timings so as to ensure a given state is reached when the chosen time delays are adversarially perturbed within a bound. In [21], we are interested in synthesizing "robust" strategies for ensuring reachability of a location in timed automata. We model this problem as a game between the controller and its environment, and solve the parameterized robust reachability problem: we show that the existence of an upper bound on the perturbations under which there is a strategy reaching a target location is EXPTIME-complete. We also extend our algorithm, with the same complexity, to turn-based timed games, where the successor state is entirely determined by the environment in some locations.

We also tackled the robustness problem for time Petri nets (TPNs, for short) in [17] by considering the model of parametric guard enlargement which allows time-intervals constraining the firing of transitions in TPNs to be enlarged by a (positive) parameter. We show that TPNs are not robust in general and checking if they are robust with respect to standard properties (such as boundedness, safety) is undecidable. We then extend the marking class timed automaton construction for TPNs to a parametric setting, and prove that it is compatible with guard enlargements. We apply this result to the (undecidable) class of TPNs which are robustly bounded (i.e., whose finite set of reachable markings remains finite under infinitesimal perturbations): we provide two decidable robustly bounded subclasses, and show that one can effectively build a timed automaton which is timed bisimilar even in presence of perturbations. This allows us to apply existing results for timed automata to these TPNs and show further robustness properties.

7.1.5. Verification for classes of Petri Nets with time

Participants: Blaise Genest, Loïc Hélouët.

We have considered verification problems for classes of Petri Nets with time. We have introduced the first, up to our knowledge, decidability result on reachability and boundedness for Petri Net variants that combine unbounded places, time, and urgency (the ability to enforce actions to happen within some delay). For this, we introduce the class of Timed-Arc Petri Nets with Urgency, which extends Timed-Arc Petri Nets [58] to allow urgency constraints, a feature from Timed-transition Petri Nets (TPNs) [54]. In order to avoid (straightforward) undecidability, we have considered restricted urgency: urgency can be used only on transitions consuming tokens from bounded places. For Timed-Arc Petri Nets with restricted urgency,

we extend decidability results from Timed-Arc Petri Nets: control-state reachability and boundedness are decidable. Our main result concerns (marking) reachability, which is undecidable for both TPNs (because of unrestricted urgency) [52] and Timed-Arc Petri Nets (because of infinite number of clocks) [57]. We have obtained decidability of reachability for (unbounded) TPNs with restricted urgency under a new, yet natural, timed-arc semantics presenting them as Timed-Arc Petri Nets with restricted urgency. Decidability of reachability under the original semantics of TPNs was also obtained for a restricted subclass of unbounded nets. This work is under submission.

7.1.6. Non-interference in partial order models

Participant: Loïc Hélouët.

In [36] we have proposed a new definition of interference for partial order models. Non-interference (NI) is a property of systems stating that confidential actions should not cause effects observable by unauthorized users. Several variants of NI have been studied for many types of models, but rarely for true concurrency or unbounded models. In [36] we have investigated NI for High-level Message Sequence Charts (HMSC), a scenario language for the description of distributed systems, based on composition of partial orders. We firstly have proposed a general definition of security properties in terms of equivalence among observations, and shown that these properties, and in particular NI are undecidable for HMSCs. We hence have considered weaker local properties, describing situations where a system is attacked by a single agent, and show that local NI is decidable in this context. We then have proposed a refinement of local NI to obtain a finer notion of causal NI that emphasizes causal dependencies between confidential actions and observations. This causal NI has then been extended to causal NI with (selective) declassification of confidential events. Finally, we have shown that checking whether a system satisfies local and causal NI and their declassified variants are PSPACE-complete problems. Decidability seems to extend to other classes of partial order models which partially ordered observations can be represented by partial order models that exhibit some forms of regularity such as graph grammars or partial order automata. This conjecture will be explored next year.

7.1.7. Synthesis and games

Participants: Ocan Sankur, Engel Lefaucheux.

In [33], we investigate compositional algorithms to solve safety games described succinctly by synchronous circuits (given by AND and inverter gates). We show how the safety specification can be decomposed, in most cases, into a set of simpler specifications, each defining a safety game depending on less inputs and state variables. We give several algorithms which consist in solving the subgames, and aggregating them in order to find strategies for the global game. We present results of extensive experiments done on around five hundred benchmarks used in the synthesis competition SYNTCOMP 2014 and show that the compositional approach improves the performence on several classes of benchmarks.

In [35] we investigate priced timed games. Priced timed games are two-player zero-sum games played on priced timed automata (whose locations and transitions are labeled by weights modeling the costs of spending time in a state and executing an action, respectively). The goals of the players are to minimise and maximise the cost to reach a target location, respectively. We consider priced timed games with one clock and arbitrary (positive and negative) weights and show that, for an important subclass (the so-called simple priced timed games), one can compute, in exponential time, the optimal values that the players can achieve, with their associated optimal strategies. As side results, we also show that one-clock priced timed games are determined and that we can use our result on simple priced timed games to solve the more general class of so-called reset-acyclic priced timed games (with arbitrary weights and one-clock).

In [34], we introduce a novel rule for synthesis of reactive systems, applicable to systems made of n components which have each their own objectives. This rule is based on the notion of admissible strategies. Intuitively, a strategy σ is dominated by σ' if against all strategies of other players, σ' is as good as σ, and against at least one strategy σ' is strictly better than σ. Admissible strategies are those that are not dominated by any other strategy. The assume-admissible synthesis consists in restricting the space of strategies to admissible ones, and to look for strategy profiles which satisfy given specifications. We compare this rule with previous

rules defined in the literature, and show that contrary to the previous proposals, it defines sets of solutions which are rectangular. This property leads to solutions which are robust and resilient, and allows one to synthesize strategies separately for each agent. We provide algorithms with optimal complexity and also an abstraction framework compatible with the new rule.

7.2. Management of large distributed systems

7.2.1. Parameterized verification in parameterized networks

Participants: Nathalie Bertrand, Paulin Fournier.

We study the problems of reaching a specific control state, or converging to a set of target states, in networks with a parameterized number of identical processes communicating via broadcast. To reflect the distributed aspect of such networks, we restrict our attention to executions in which all the processes must follow the same local strategy that, given their past performed actions and received messages, provides the next action to be performed. We show that the reachability and target problem under such local strategies are NP-complete, assuming that the set of receivers is chosen non-deterministically at each step. On the other hand, these problems become undecidable when the communication topology is a clique. However, decidability can be regained with the additional assumption that all processes are bound to receive the broadcast messages. This is a joint work with Arnaud Sangnier [31].

7.2.2. Runtime enforcement of untimed and timed properties

Participants: Thierry Jéron, Hervé Marchand, Srinivas Pinisetty.

Runtime enforcement is a powerful technique to ensure that a running system satisfies some desired properties. Using an enforcement monitor, an (untrustworthy) input execution (in the form of a sequence of events) is modified into an output sequence that complies with a property. Over the last decade, runtime enforcement has been mainly studied in the context of untimed properties. For several years, and in particular in the context of the PhD thesis of Srinivas Pinisetty [15] we elaborated the theory of runtime enforcement of timed properties. This year we also continued our work on the subject in several directions.

In [38] we describe the TiPEX tool that implements the enforcement monitoring algorithms for timed properties proposed in our previous papers . Enforcement monitors are generated from timed automata specifying timed properties. Such monitors correct input sequences by adding extra delays between events. Moreover, TiPEX also provides modules to generate timed automata from patterns, compose them, and check the class of properties they belong to in order to optimize the monitors. This paper also presents the performance evaluation of TiPEX within some experimental setup.

With coleagues from LaBRI (M. Renard, A. Rollet) and LIG (Y. Falcone) we investigate runtime enforcement of (timed and untimes) properties with uncontrollable events. In [41], we introduce a framework that takes as input any regular (timed) property over an alphabet of events, with some of these events being uncontrollable. An uncontrollable event cannot be delayed nor intercepted by an enforcement mechanism. Enforcement mechanisms satisfy important properties, namely soundness and compliance, meaning that enforcement mechanisms output correct executions that are close to the input execution. We discuss the conditions for a property to be enforceable with uncontrollable events, and we define enforcement mechanisms that modify executions to obtain a correct output, as soon as possible. Moreover, we synthesize sound and compliant descriptions of runtime enforcement mechanisms at two levels of abstraction to facilitate their design and implementation.

With colleagues from the Aalto University (S. Pinisetty, S. Tripakis and V. Preoteasa) and LIG (Y. Falcone) we investigate predictive runtime enforcement. In [39] we introduce predictive runtime enforcement, where the system is not entirely black-box, but we know something about its behavior. This a-priori knowledge about the system allows to output some events immediately, instead of delaying them until more events are observed, or even blocking them permanently. This in turn results in better enforcement policies. We also show that if we have no knowledge about the system, then the proposed enforcement mechanism reduces to a classical non-predictive RE framework. All our results are formalized and proved in the Isabelle theorem prover. We are also currently extending this work to the timed setting.

7.2.3. Discrete controller synthesis

Participants: Nicolas Berthier, Hervé Marchand.

In [29] we investigate the opportunities given by recent developments in the context of Discrete Controller Synthesis algorithms for infinite, logico-numerical systems. To this end, we focus on models employed in previous work for the management of dynamically partially reconfigurable hardware architectures. We extend these models with logico-numerical features to illustrate new modeling possibilities, and carry out some benchmarks to evaluate the feasibility of the approach on such models.

In [30] we elaborate on our former work for the safety control of infinite reactive synchronous systems modeled by arithmetic symbolic transition systems. By using abstract interpretation techniques involving disjunctive polyhedral overapproximations, we provide effective symbolic algorithms allowing to solve the deadlock-free safety control problem while overcoming previous limitations regarding the non-convexity of the set of states violating the invariant to enforce.

The ever growing complexity of software systems has led to the emergence of automated solutions for their management. The software assigned to this work is usually called an Autonomic Management System (AMS). It is ordinarily designed as a composition of several managers, which are pieces of software evaluating the dynamics of the system under management through measurements (e.g., workload, memory usage), taking decisions, and acting upon it so that it stays in a set of acceptable operating states. However, careless combination of managers may lead to inconsistencies in the taken decisions, and classical approaches dealing with these coordination problems often rely on intricate and ad hoc solutions. To tackle this problem, we take a global view and underscore that AMSs are intrinsically reactive, as they react to flows of monitoring data by emitting flows of reconfiguration actions. Therefore in [19] we propose a new approach for the design of AMSs, based on synchronous programming and discrete controller synthesis techniques. They provide us with high-level languages for modeling the system to manage, as well as means for statically guaranteeing the absence of logical coordination problems. Hence, they suit our main contribution, which is to obtain guarantees at design time about the absence of logical inconsistencies in the taken decisions. We detail our approach, illustrate it by designing an AMS for a realistic multi-tier application, and evaluate its practicality with an implementation.

In the invited paper [24] we make an overview of our works addressing discrete control-based design of adaptive and reconfigurable computing systems, also called autonomic computing. They are characterized by their ability to switch between different execution modes w.r.t. application and functionality, mapping and deployment, or execution architecture. The control of such reconfigurations or adaptations is a new application domain for control theory, called feedback computing. We approach the problem with a programming language supported approach, based on synchronous languages and discrete control synthesis. We concretely use this approach in FPGA-based reconfigurable architectures, and in the coordination of administration loops.

7.2.4. Computing knowledge at runtime

Participant: Blaise Genest.

In [37] we compare three notions of knowledge in concurrent system: memoryless knowledge, knowledge of perfect recall, and causal knowledge. Memoryless knowledge is based only on the current state of a process, knowledge of perfect recall can take into account the local history of a process, and causal knowledge depends on the causal past of a process, which comprises the information a process can obtain when all processes exchange the information they have when performing joint transitions. We compare these notions in terms of knowledge strength, number of bits required to store this information, and the complexity of checking if a given process has a given knowledge. We show that all three notions of knowledge can be implemented using finite memory. Causal knowledge proves to be strictly more powerful than knowledge with perfect recall, which in turn proves to be strictly more powerful than memoryless knowledge. We show that keeping track of causal knowledge is cheaper than keeping track of knowledge of perfect recall.

7.2.5. Distributed optimal 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 local plans, one per component, such that their combination forms a consistent global plan of minimal cost.

Our previous solutions to this problem modeled components as weighted automata [22]. In collaboration with Loig Jezequel (TU Munich) and Victor Khomenko (Univ. of Newcastle), we have extended this approach to the case of components modeled as safe Petri nets[23]. 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 (our work also contains a contribution to this difficult question).

7.2.6. Regulation of urban train systems

Participants: Éric Fabre, Loïc Hélouët, Karim Kecir, Hervé Marchand, Christophe Morvan.

A part of the SUMO team is involved in a collaboration with Alstom transports on regulation techniques. The role of regulation algorithms is to observe train trajectories and delays with respect to an expected ideal schedule, and then compute commands that are sent to trains to meet some quality of service (punctuality, regularity, ...) The objective of this collaboration is to study regulation techniques that are currently in use in urban train systems and compare their performances, and in the future to be able to compute optimal regulation strategies.

This year, we have proposed models inspired from stochastic Petri nets and from closed loop controllers to simulate regulated railways systems. The Petri net model led to the design of a tool called SIMSTORS, that was sucessfully used to model a real case study (line 1 of Santiago's subway). The simulator relies on event-based symbolic techniques: the time elapsed between two steps of the simulation is the time between two event occurrences (arrival, departure of a train, incident,...). This simulation scheme relying on an abstract model allowed a dramatic speed up of simulation with respect to existing solutions in use at Alstom Transport.

A second line of work has also been explored, in order to design and evaluate new regulation strategies for subway lines. The underlying model is inspired from event-based control theory, in a stochastic and timed setting. It abstracts away several significant topological features of a subway line, and focuses on the optimal command of train speeds in order to achieve high-level objectives such as the equal spacing of trains, or the efficient insertion/extraction of trains. This approach has allowed us to design new distributed regulation policies, which are remarquably stable and efficiently mitigate known instabilities of subway lines, like the bunching phenomenon. We are currently working on an extension of this approach for the management of time-tables and of forks and joins in the topology of subway lines.

7.3. Data driven systems

7.3.1. A model of large-scale distributed collaborative system

Participants: Éric Badouel, Loïc Hélouët, Christophe Morvan, Robert Nsaibirni.

We have presented in [27] and [18] a purely declarative approach to artifact-centric collaborative systems, a model which we introduced in two stages. First, we assume that the workspace of a user is given by a mindmap, shortened to a map, which is a tree used to visualize and organize tasks in which he or she is involved, together with the information used for the resolution of these tasks. We introduce a model of guarded attribute grammar, or GAG, to help the automation of updating such a map. A GAG consists of an underlying grammar, that specifies the logical structure of the map, with semantic rules which are used both to govern the evolution of the tree structure (how an open node may be refined to a subtree) and to compute the value of some of the attributes (which derives from con-textual information). The map enriched with this extra information

is termed an active workspace. Second, we define collaborative systems by making the various user's active workspaces communicate with each other. The communication uses message passing without shared memory thus enabling convenient distribution on an asynchronous architecture. A case study on a disease surveillance system is under development in the PhD thesis of Robert Nsaibirni and a first prototype of the model of active workspaces was written by Eric Badouel.

7.3.2. Petri Nets with semi-structured data

Participants: Éric Badouel, Loïc Hélouët, Christophe Morvan.

In [28], we have proposed an extension of Petri nets with data called Structured Data Nets (StDN). This extension allows for the description of transactional systems with data. In StDNs, tokens are structured documents. Each transition is attached to a query, guarded by patterns, (logical assertions on the contents of its preset) and transforms tokens. In [28], we have proposed a semantics for StDNs, and then considered their formal properties: coverability of a marking, termination and soundness of transactions. Unrestricted StDNs are Turing complete, so these properties are undecidable. However, we have proposed an order on structured documents, and shown that under reasonable restrictions on documents and on the ex- pressiveness of patterns and queries, StDNs are well-structured transition systems, for which coverability, termination and soundness are decidable. This work has then been extended to consider properties of sets of configurations described as upward closed sets satisfying patterns, and should appear in a journal paper in 2016.

TASC Project-Team

7. New Results

7.1. IBEX

The development of the Ibex library has continued. The main developments in 2015 are:

- the complete refactoring of the multi-heap internal structure used for search space exploration in the global optimizer
- the creation of a new module for explicit set (or pavings) manipulation/algebra with full documentation and tutorial

7.2. NetWMS2

New advances have been made in the context of packing curved objects. The packing algorithm developed in 2014 have been published in ICJAI'15, along with new features. The calculation of the *penetration depth* (a classical measure of violation cost for overlapping objects) has also been extended to the case of parametric curves (like, e.g., Bezier curves) and new experiments have been conducted with our solver for this new type of objects.

We deal with the problem of packing two-dimensional objects of quite arbitrary shapes including in particular curved shapes (like ellipses) and assemblies of them. This problem arises in industry for the packaging and transport of bulky objects which are not individually packed into boxes, like car spare parts. There has been considerable work on packing curved objects but, most of the time, with specific shapes; one famous example being the circle packing problem. There is much less algorithm for the general case where different shapes can be mixed together. A successful approach has been proposed recently by Martinez et al. and the algorithm we propose here is an extension of their work. Martinez et al. use a stochastic optimization algorithm with a fitness function that gives a violation cost and equals zero when objects are all packed. Their main idea is to define this function as a sum of $\binom{n}{2}$ elementary functions that measure the overlapping between each pair of different objects. However, these functions are ad-hoc formulas. Designing ad-hoc formulas for every possible combination of object shapes can be a very tedious task, which dramatically limits the applicability of their approach. We generalize the approach by replacing the ad-hoc formulas with a numerical algorithm that automatically measures the overlapping between two objects. Then, we come up with a fully black-box packing algorithm that accept any kind of objects.

7.3. Time-Series Constraints

We describe a large family of constraints for structural time series by means of function composition. These constraints are on aggregations of features of patterns that occur in a time series, such as the number of its peaks, or the range of its steepest ascent. The patterns and features are usually linked to physical properties of the time series generator, which are important to capture in a constraint model of the system, i.e. a conjunction of constraints that produces similar time series. We formalise the patterns using finite transducers, whose output alphabet corresponds to semantic values that precisely describe the steps for identifying the occurrences of a pattern. Based on that description, we automatically synthesise automata with accumulators, as well as constraint checkers. The description scheme not only unifies the structure of the existing 30 time-series constraints in the Global Constraint Catalogue, but also leads to over 600 new constraints, with more than 100,000 lines of synthesised code.

7.4. New Global Constraints

This year we introduce new generic global constraints that can be respectively used to reformulate a number of constraints where the formulation become easy once some tuples are sorted, and to express temporal relation between two sequence of intervals.

- Some constraint programming solvers and constraint modelling languages feature the *sort*(L, P, S) constraint, which holds if S is a nondecreasing rearrangement of the list L, the permutation being made explicit by the optional list P. However, such sortedness constraints do not seem to be used much in practice. We argue that reasons for this neglect are that it is impossible to require the underlying sort to be stable, so that *sort* cannot be guaranteed to be a total-function constraint, and that L cannot contain tuples of variables, some of which form the key for the sort. To overcome these limitations, we introduce the *stable-keysort* constraint, decompose it using existing constraints, and propose a propagator. This new constraint enables a powerful modelling idiom, which we illustrate by elegant and scalable models of two problems that are otherwise hard to encode as constraint programs.
- The constraint was initially motivated by an application where the objective is to generate a video summary, built using intervals extracted from a video source. In this application, the constraints used to select the relevant pieces of intervals are based on Allen's algebra. The best state-of-the-art results are obtained with a small set of ad hoc solution techniques, each specific to one combination of the 13 Allen's relations. Such techniques require some expertise in Constraint Programming. This is a critical issue for video specialists. We design a generic constraint, dedicated to a class of temporal problems that covers this case study, among others. ExistAllen takes as arguments a vector of tasks, a set of disjoint intervals and any of the 213 combinations of Allen's relations. ExistAllen holds if and only if the tasks are ordered according to their indexes and for any task at least one relation is satisfied, between the task and at least one interval. We design a propagator that achieves bound-consistency in O(n+m), where n is the number of tasks and m the number of intervals. This propagator is suited to any combination of Allen's relations, without any specific tuning. Therefore, using our framework does not require a strong expertise in Constraint Programming. The experiments, performed on real data, confirm the relevance of our approach.

7.5. Controlling the Generation of Solutions

The following two results deal with controlling the generation of solutions to a constraint problem.

- The *focus* constraint expresses the notion that solutions are concentrated. In practice, this constraint suffers from the rigidity of its semantics. To tackle this issue, we propose three generalizations of the FOCUS constraint. We provide for each one a complete filtering algorithm. Moreover, we propose mathematical programming (ILP) and constraint programming decompositions.
- There are significant motivations for considering alternate solutions to a problem. As expressed by renowned statistician George Box *The most that can be expected from any model is that it can supply a useful approximation to reality: all models are wrong; some models are useful.*. Multiple solutions alone, however, are not sufficient to guarantee anything of value. If they are nearly identical nothing is gained. While most frameworks in the literature consider diversity between solutions through mathematical distances, this paper proposes alternative distance measures represented by global constraints. It introduces a constraint programming framework for optimization problems, able to generate sets of nearly-optimal solutions that are diverse. With respect to over-constrained problems, the framework can be specialized in order to generate solution sets where constraint violations are diverse.

TEA Project-Team

7. New Results

7.1. Polychronous automata

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

We have defined a model of *polychronous automata* based on clock relations [13]. A specificity of this model is that an automaton is submitted to clock constraints: these finite-state automata define transition systems to express explicit reactions together with properties, in the form of Boolean formulas over logical time, to constrain their behavior. This allows one to specify a wide range of control-related configurations, either reactive, or restrictive with respect to their control environment. A semantic model is defined for these polychronous automata, that relies on a Boolean algebra of clocks. Polychronous automata integrate smoothly with data-flow equations in the polychronous model of computation.

This formal model of automata also supports the recommendations adopted by the SAE committee on the AADL to implement a timed and synchronous behavioural annex for the standard [0].

A minimal syntactic extension of the Signal language has been defined to integrate polychronous automata in Polychrony. We have added a new syntactic category of *process*, called `automaton`. In such an automaton process, labeled processes represent states, and generic processes such as `Transition` are used to represent the automaton features. Usual equations can be used in these automaton processes to specify constraints or to define computations.

We have also defined and implemented the refinement of Signal processes as automata. A given Signal program may be seen as an automaton which contains one single state and one single transition, labeled by a clock. This clock is the upper bound of all the clocks of the program (the *tick* of the program). The construction of a refined automaton from a Signal program is based on delayed signals, viewed as state variables (in particular Boolean ones). A state of the automaton is a Signal program with some valuation of its state variables. Transitions are labeled by clocks, which represent the events that fire these transitions. The principle of the construction consists in dividing a given state according to the possible values of a state variable (i.e., *true* and *false* for Boolean state variables) in order to get two states, and thus two new Signal programs. Each one of these two states is obtained using a rewriting of the starting program. Moreover, the absence of value for the state variable (which can be considered as another possible value) is taken into account in the clocks labelling the transitions. The construction of the automaton is a hierarchic process. Thanks to the clock hierarchy, this construction, which would be expensive in the worst case (the size of the explicit automaton being an exponential of its number of state variables), may be heavily simplified.

7.2. Runtime verification and trace analysis

Participants: Vania Joloboff, Daian Yue, Frédéric Mallet.

When engineers design a new cyber physical system, there are well known requirements that can be translated as system properties that must be verified. These properties can be expressed in some formalism and when the model has been designed, the properties can be checked at the model level, using model checking techniques or other model verification techniques. When building a virtual prototype of the system, including a combination of simulated hardware, firmware and application software, the executable models can be augmented also with property verification, for example in the PSL language, or simply by introducing assertions in the implementation code.

[0] *Logically timed specification in the AADL: a synchronous model of computation and communication (recommendations to the SAE committee on AADL).* L. Besnard, E. Borde, P. Dissaux, T. Gautier, P. Le Guernic, and J.-P. Talpin. Technical Report RT-0446, Inria, 2014.

This requires that the properties are well specified at the time the virtual prototype is assembled. However it is also the case that many intrinsic properties are actually unforeseen when the virtual prototype is assembled, for example that some hardware buffer overflow should not remain unnoticed by the software. In most cases, during system design the simulation fails: the engineers then must investigate the cause of the failure. Most of time the failure is due to an unexpected sequence of states and transitions that involve several components mixing hardware and software that could not be checked at the model level (e.g. state explosion) or was simply unforeseen. The engineers then have to investigate the cause of failure.

A widely used technique for that consists in storing all of the trace data of simulation sessions into trace files, which are analyzed later with specialized trace analyzer tools. Such trace files have become huge, possibly hundred of Gigabytes as all data are stored into the trace files, and have become untractable by human manual handling. The engineers use some kind of search tools to identify the cause of failure and after iterative refinement steps, which are very time consuming, eventually identify the reason, most often some unforeseen causality chain of events and state transitions that lead to a failure. A new system property can then be captured and included into the set of verified properties.

In order to better identify the reason for such failures and capture the missing properties that the system should verify we have started to work on a new run time verification approach based on trace analysis. Approaches like PSL requires that the properties to verify are known before hand. Our approach is attempting for the engineers to experiment various property verification of failing simulations without re-building the virtual prototype. We are investigating a technique for trace analysis that makes it possible to investigate properties either statically working from a trace file or dynamically by introducing a dynamic verification component into the virtual prototype.

The first idea is to introduce a formal mapping/filtering technique such that the raw data generated by a virtual prototype can be mapped onto a formal trace model. For that, we propose to use a model transformer whose code is generated from a higher level. Using the Eclipse modelling framework, we propose for the virtual prototyping engineers to first describe using a Domain Specific Language how the raw output of the simulator can be filtered and mapped to a formal model. This Domain Specific Language takes as input the description of the simulator output, and the description of the formal output, following fixed meta models. In current version, the meta model of the virtual prototype dictates that it generates 'trace items' where each trace item is specified as a sequence of identified binary data variables (bits, bytes, words..) that carry a timestamp.

The model transformer generates code (in our case C++) that is dynamically invoked by the virtual prototype to dynamical map the trace output. An advantage of doing that is that all irrelevant data with regards to a tested property can be ignored and the size of trace files can be considerably reduced. For our experiment, we have chosen logical clock CCSL as our formal target formalism. The Eclipse EMF tool we have defined allows users to define a mapping model from the local simulation events from the SimSoC simulator to a logical clock format.

The second idea is to hide the complexity of the formal method formulas into a user friendly property specification language. For example, we do not want to expose the end-users engineers to understand the intricacies of CCSL or LTL. The property specification language is translated into CCSL formulas, which in turn generate automata. It should be possible then, to some extent, to change the formalism underneath the language without changing the properties expressed by the user.

The property specification language ultimately compiles into automata that parse the formal trace output generated above. At runtime of the virtual prototype, the mapping library is dynamically loaded by the simulator and generates input for the automata. The verification of the properties can be dynamic, with a true runtime verification, or statically by analyzing the (much smaller) trace file after a failure.

This year we have investigated this approach, designed the architecture described above and carried some experimental work, but a significant part of the implementation still remains to be done. We have started designing a new property specification language where the users can express properties such as causality (e.g. the train must not start if the door is opened) or jittering or clock drift in image processing [11], [10]. There remain some theoretical issues with regards to which properties can be effectively verified.

7.3. Integration of Polychrony with QGen

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

The FUI project P gave birth to the QGen qualifiable model compiler, developed by Adacore. The tool accepts a discrete subset of Simulink expressed in a language called P and produces C or Ada code. It is currently not known if an architectural description language is going to be integrated in QGen, as originally planned.

We developed a transformation tool named `P2S` for expressing P system models in Signal, using the EMF (Eclipse Modelling Framework) technology. `P2S` tool is written in Clojure, a language inspired by Lisp running on the Java Virtual Machine, which helped us define a terse and expressive API for manipulating Signal models while remaining fully interoperable with existing Java libraries (including Eclipse plugins and especially Polychrony ones).

We experimented this transformation tool on small to medium use cases provided by members of the P project. Our work is detailed in a conference paper titled "Integration of Polychrony and QGen Model Compiler", which will appear at ERTSS'16 [0]. A perspective of our work is to convert the intermediate code emitted by QGen as Signal too (under development), in order to produce a fully executable Signal model of Simulink models, and combine them with architectural description of systems in AADL, and/or P's architecture language.

7.4. Formal semantics and model-based analysis of AADL specifications

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

Last year, the SAE committee on the AADL adopted our recommendations to implement a timed and synchronous behavioural annex for the standard. We have defined a new model of polychronous constrained automata that has been provided as semantic model for our proposal of an extension of the AADL behavioural annex. An experimental implementation of the semantic features of this "timing annex" will be provided through the Polychrony framework. For that purpose, representations of automata have been introduced in the Signal toolbox of Polychrony. The implementation will enrich the already existing transformation from AADL models to Signal programs to consider behaviour of AADL models, and will be integrated in the POP environment for Eclipse. The transformation from AADL behaviour annex to Signal programs use the Signal extension for polychronous automata, which are used as the common semantic domain. The implementation is currently tested with the adaptive cruise control case study developed with Toyota ITC.

Our work with the SAE committee is sponsored by Toyota, with whom we started a new project in 2014 jointly with VTRL as US partner. The main topic of our project is the semantic-based model integration of automotive architectures, virtual integration, toward formal verification and automated code synthesis [19]. The project led to the elaboration of a case study of an adaptive cruise control system, supported through an AADL implementation and a video of demonstration. The case study implementation is an AADL model representing the whole adaptive cruise control system, from car devices (e.g., brakes, throttle or radar) to software behavior, including embedded hardware (buses, processors and memories). It will be used in the future to demonstrate property and constraint analyses through heterogeneous systems. Huafeng Yu, our main collaborator at Toyota ITC, presented the video of demonstration at the annual Toyota show case. Early returns from the show case express a growing interest of Toyota for architecture and timing of car embedded systems, which could lead to new collaborations.

7.5. Refinement types for reactive system models

Participants: Pierre Jouvelot, Sandeep Shukla, Jean-Pierre Talpin.

[0] *Integration of polychrony in the QGen model compiler*. C. Junke, T. Gautier, L. Besnard, J.-P. Talpin. ERTS'16 - European Congress on Embeddd Real-Rime Software and Systems, 2016.

We introduced a new technique born from the field of functional programming to adapt and extend it to the case of reaction systems, the notion of refinement types of Jahla et al. [0]. Our idea is to formulate the analysis of algebraic properties in synchronous and reactive programs as data-dependent type properties formulated using multi-sorted logic formulas, which we call liquid clocks [20], [18]. Our objectives are to cover the case of several models of concurrency and computation: synchronous, asynchronous, data-parallel; as well as to formulate such algebraic properties for linear, continuous and logical forms of time, all into the same type-theoretical framework. This work, born from two collaborations With USAF/VT and with the ANR Feever project, will be pursued within the TIX international partnership.

7.6. Formal verification of timing aspects of cyber-physical systems using a contract theory

Participants: Jean-Pierre Talpin, Benoit Boyer, David Mentre, Simon Lunel.

This is a new project in collaboration with Mitsubishi Electronics Research Centre Europe (MERCE). The primary goal of our project is to ensure correctness-by-design in cyber-physical systems, i.e., systems that mix software and hardware in a physical environment, e.g., Mitsubishi factory automation lines. We plan to explore a multi-sorted algebraic framework for static analysis and formal verification starting from a simple use case extracted from Mitsubishi factory automation documentations. This will serve as a basis to more ambitious research where we intend to leverage recent advance in type theory, SMT solvers for nonlinear real arithmetic (dReal and δ-decidability) and contracts theory (meta-theory of Benveniste et al., Ruchkin's contracts) to provide a general framework of reasoning about heterogeneous factory components.

[0] *Liquid Types*. P. M. Rondon, M. Kawaguchi, R. Jhala. PLDI, 2008

ASPI Project-Team

5. New Results

5.1. Adaptive multilevel splitting

Participants: Frédéric Cérou, Arnaud Guyader.

We have show last year that an adaptive version of multilevel splitting for rare events is strongly consistent and 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. This year we have generalized these results to include Markov kernels used to move the particles (or *shakers*) are of Metropolis–Hastings type. This is a non–trivial generalization to a very important case.

5.2. Adaptive multilevel splitting as a Fleming–Viot system

Participants: Frédéric Cérou, Arnaud Guyader.

This is a collaboration with Bernard Delyon (université de Rennes 1) and Mathias Rousset (EPI MATHERIALS, Inria Paris Rocquencourt).

By considering the adaptive multilevel splitting algorithm as a Fleming–Viot particle system for a stochastic wave, in the sense of [42], we have shown the mean square convergence using a general result [67] about the convergence of Fleming–Viot (Villemonais, 2013). We are currently working on the proof of a central limit theorem, but the proof is not yet complete. We have nevertheless identified the expression of the asymptotic variance.

5.3. Bias and variance reduction in rare event simulation

Participant: François Le Gland.

This is a collaboration with Damien Jacquemart (ONERA, Palaiseau) and Jérôme Morio (ONERA, Toulouse).

In [17], we highlight a bias induced by the discretization of the sampled Markov paths in the splitting algorithm, and we propose to correct this bias using a deformation of the intermediate regions, as proposed in [48]. Moreover, we propose two numerical methods to design intermediate regions in the splitting algorithm that minimise the variance. One is connected with a partial differential equation approach, the other one is based on the discretization of the state space of the process.

5.4. Simulation–based algorithms for the optimization of sensor deployment

Participant: François Le Gland.

This is a collaboration with Christian Musso (ONERA, Palaiseau) and with Sébastien Paris (LSIS, université du Sud Toulon Var).

The problem considered here can be described as follows: a limited number of sensors should be deployed by a carrier in a given area, and should be activated at a limited number of time instants within a given time period, so as to maximize the probability of detecting a target (present in the given area during the given time period). There is an information dissymmetry in the problem: if the target is sufficiently close to a sensor position when it is activated, then the target can learn about the presence and exact position of the sensor, and can temporarily modify its trajectory so as to escape away before it is detected. This is referred to as the target intelligence. Two different simulation–based algorithms have been designed in [23] to solve separately or jointly this optimization problem, with different and complementary features. One is fast, and sequential: it proceeds by running a population of targets and by dropping and activating a new sensor (or re–activating a sensor already available) where and when this action seems appropriate. The other is slow, iterative, and non–sequential: it proceeds by updating a population of deployment plans with guaranteed and increasing criterion value at each iteration, and for each given deployment plan, there is a population of targets running to evaluate the criterion. Finally, the two algorithms can cooperate in many different ways, to try and get the best of both approaches. A simple and efficient way is to use the deployment plans provided by the sequential algorithm as the initial population for the iterative algorithm.

5.5. Kalman Laplace filtering

Participant: François Le Gland.

This is a collaboration with Paul Bui Quang (CEA, Bruyères–le–Châtel) and Christian Musso (ONERA, Palaiseau).

We propose in [21] a new nonlinear Bayesian filtering algorithm where the prediction step is performed like in the extended Kalman filter, and the update step is done thanks to the Laplace method for integral approximation. This algorithm is called the Kalman Laplace filter (KLF). The KLF provides a closed–form non–Gaussian approximation of the posterior density. The hidden state is estimated by the maximum a posteriori. We describe a way to alleviate the computation cost of this maximization, when the likelihood is a function of a vector whose dimension is smaller than the state space dimension. The KLF is tested on three simulated nonlinear filtering problems: target tracking with angle measurements, population dynamics monitoring, motion reconstruction by neural decoding. It exhibits a good performance, especially when the observation noise is small.

5.6. Combining analog method and ensemble data assimilation

Participants: François Le Gland, Valérie Monbet, Chau Thi Tuyet Trang.

This is a collaboration with Pierre Ailliot (université de Bretagne Occidentale), Ronan Fablet and Pierre Tandéo (Télécom Bretagne), Anne Cuzol (université de Bretagne Sud) and Bernard Chapron (IFREMER, Brest).

Nowadays, ocean and atmosphere sciences face a deluge of data from spatial observations, in situ monitoring as well as numerical simulations. The availability of these different data sources offer new opportunities, still largely underexploited, to improve the understanding, modeling and reconstruction of geophysical dynamics. The classical way to reconstruct the space–time variations of a geophysical system from observations relies on data assimilation methods using multiple runs of the known dynamical model. This classical framework may have severe limitations including its computational cost, the lack of adequacy of the model with observed data, modeling uncertainties. In [24], we explore an alternative approach and develop a fully data–driven framework, which combines machine learning and statistical sampling to simulate the dynamics of complex system. As a proof concept, we address the assimilation of the chaotic Lorenz–63 model. We demonstrate that a nonparametric sampler from a catalog of historical datasets, namely a nearest neighbor or analog sampler, combined with a classical stochastic data assimilation scheme, the ensemble Kalman filter and smoother, reach state–of–the–art performances, without online evaluations of the physical model.

5.7. Markov–switching vector autoregressive models

Participant: Valérie Monbet.

This is a collaboration with Pierre Ailliot (université de Bretagne Occidentale), Julie Bessac (Argonne National Laboratory, Chicago) and Julien Cattiaux (Météo–France, Toulouse).

Multivariate time series are of interest in many fields including economics and environment. The most popular tools for studying multivariate time series are the vector autoregressive (VAR) models because of their simple specification and the existence of efficient methods to fit these models. However, the VAR models do not allow to describe time series mixing different dynamics. For instance, when meteorological variables are observed, the resulting time series exhibit an alternance of different temporal dynamics corresponding to weather regimes. The regime is often not observed directly and is thus introduced as a latent process in time series models in the spirit of hidden Markov models. Markov switching vector autoregressive (MSVAR) models have been introduced as a generalization of autoregressive models and hidden Markov models. They lead to flexible and interpretable models. In this mutivariate context, several questions occur.

- The discrete hidden variable also called regime has to be correctly defined. Indeed the regime can be local (e.g. link to a subset of the variables) or global (e.g. the same for all the variables). It can also be observed and inferred a priori or hidden. In the second case, it has to be estimated at the same time as the model parameters.
  The question of the definition of the regime is investigated in [26] for the specific problem of multi site wind modeling.
- Markov Switching VAR models (MSVAR) suffer of the same dimensionality problem as VAR models. For large (and even moderate) dimensions, the number of autoregressive coefficients in each regime can be prohibitively large which results in noisy estimates. When the variables are correlated, which is the standard situation in multivariate time series, over–learning is frequent. The estimated parameters contains spurious non–zero coefficients and are then difficult to interpret. The predictions associated to the model are usually unstable. Collinearity causes also ill–conditioning of the innovation covariance. In [29], we propose a likelihood penalization method with hard thresholding for MSVAR models leading to sparse MSVAR. Both autoregressive matrices and precision matrices are penalized using smoothly clipped absolute deviation (SCAD) penalties.

5.8. Dependent time changed processes

Participant: Valérie Monbet.

This is a collaboration with Pierre Ailliot (université de Bretagne Occidentale), Bernard Delyon (université de Rennes 1) and Marc Prevosto (IFREMER, Brest).

Many records in environmental sciences exhibit asymmetric trajectories and there is a need for simple and tractable models which can reproduce such feature. In [25] we explore an approach based on applying both a time change and a marginal transformation on Gaussian processes. The main originality of the proposed model is that the time change depends on the observed trajectory. We first show that the proposed model is stationary and ergodic and provide an explicit characterization of the stationary distribution. This result is then used to build both parametric and non–parametric estimate of the time change function whereas the estimation of the marginal transformation is based on up–crossings. Simulation results are provided to assess the quality of the estimates. The model is applied to wave data and it is shown that the fitted model is able to reproduce important statistics of the data such as its spectrum and marginal distribution which are important quantities for practical applications. An important benefit of the proposed model is its ability to reproduce the observed asymmetries between the crest and the troughs and between the front and the back of the waves by accelerating the chronometer in the crests and in the front of the waves.

5.9. An efficient algorithm for video super–resolution based on a sequential model

Participant: Patrick Héas.

This is a collaboration with Angélique Drémeau (ENSTA Bretagne, Brest) and Cédric Herzet (EPI FLUMINANCE, Inria Rennes–Bretagne Atlantique)

In the work [27], we propose a novel procedure for video super–resolution, that is the recovery of a sequence of high–resolution images from its low–resolution counterpart. Our approach is based on a *sequential* model (i.e. each high–resolution frame is supposed to be a displaced version of the preceding one) and considers the use of sparsity–enforcing priors. Both the recovery of the high–resolution images and the motion fields relating them is tackled. This leads to a large–dimensional, non–convex and non–smooth problem. We propose an algorithmic framework to address the latter. Our approach relies on fast gradient evaluation methods and modern optimization techniques for non–differentiable/non–convex problems. Unlike some other previous works, we show that there exists a provably–convergent method with a complexity linear in the problem dimensions. We assess the proposed optimization method on several video benchmarks and emphasize its good performance with respect to the state of the art.

5.10. Reduced–order modeling of hidden dynamics

Participant: Patrick Héas.

This is a collaboration with Cédric Herzet (EPI FLUMINANCE, Inria Rennes–Bretagne Atlantique).

The objective of the paper [28] is to investigate how noisy and incomplete observations can be integrated in the process of building a reduced–order model. This problematic arises in many scientific domains where there exists a need for accurate low–order descriptions of highly–complex phenomena, which can not be directly and/or deterministically observed. Within this context, the paper proposes a probabilistic framework for the construction of POD–Galerkin reduced–order models. Assuming a hidden Markov chain, the inference integrates the uncertainty of the hidden states relying on their posterior distribution. Simulations show the benefits obtained by exploiting the proposed framework.

I4S Project-Team

7. New Results

7.1. Reflectometry

7.1.1. Experimental validation of the inverse scattering method for distributed characteristic impedance estimation

Participant: Qinghua Zhang.

This work has been carried out in collaboration with Florent Loete (GEEPS-SUPELEC) and with Michel Sorine, formerly member of the Inria SISYPHE EPI.

Recently published theoretic results and numerical simulations have shown the ability of inverse scattering-based methods to diagnose soft faults in electric cables, in particular, faults implying smooth spatial variations of cable characteristic parameters. The purpose of the present work is to realize laboratory experiments confirming the ability of the inverse scattering method for retrieving spatially distributed characteristic impedance from reflectometry measurements. Various smooth or stepped spatial variations of characteristic impedance profiles are tested. This study has been accomplished in the framework of the ANR SODDA project and the results have been published in IEEE Transactions on Antennas and Propagation [16].

7.2. Automatic control

7.2.1. Observability conservation by output feedback and observability Gramian bounds

Participants: Qinghua Zhang, Liangquan Zhang.

Though it is a trivial fact that the observability of a linear state space system is conserved by output feedback, it requires a rigorous proof to generalize this result to uniform complete observability, which is defined with the observability Gramian. The purpose of this work is to complete such a proof. Some issues in existing results are also discussed. The uniform complete observability of closed loop systems is useful for the analysis of some adaptive systems and of the Kalman filter. This study has been accomplished in the framework of the ITEA MODRIO project and the results have been published in Automatica [20].

7.2.2. Weighted principal component analysis for Wiener system Identification: regularization and non-Gaussian excitations

Participant: Qinghua Zhang.

This work has been carried out in collaboration with Vincent Laurain (CRAN/CNRS/Université de Lorraine) and with Jiandong Wang (Peking University).

Finite impulse response (FIR) Wiener systems driven by Gaussian inputs can be efficiently identified by a well-known correlation-based method, except those involving even static nonlinearities. To overcome this deficiency, another method based on weighted principal component analysis (wPCA) has been recently proposed. Like the correlation-based method, the wPCA is designed to estimate the linear dynamic subsystem of a Wiener system without assuming any parametric form of the nonlinearity. To enlarge the applicability of this method, it is shown in this work that high order FIR approximation of IIR Wiener systems can be efficiently estimated by controlling the variance of parameter estimates with regularization techniques. The case of non-Gaussian inputs is also studied by means of importance sampling. The results of this study have been presented in [22].

7.2.3. LPV system common state basis estimation from independent local LTI models

Participant: Qinghua Zhang.

This work has been carried out in collaboration with Lennart Ljung (Linköping University).

For the identification of a linear parameter varying (LPV) system steered by a scheduling variable evolving within a finite set, the local approach consists in separately estimating local linear time invariant (LTI) models corresponding to fixed values of the scheduling variable. It is shown in this work that, without any global structural assumption of the considered LPV system, the local state-space LTI models do not contain the necessary information about the similarity transformations making them coherent. Nevertheless, it is possible to estimate these similarity transformations from input-output data under appropriate input excitation conditions. These estimations result in a common state basis of the transformed local LTI models, so that they form a coherent global LPV model, suitable for numerical simulations in the case of fast scheduling variable evolutions. This study has been accomplished in the framework of the ITEA MODRIO project and the results have been presented in [39].

7.3. Damage detection and linear state analysis

7.3.1. Vibration monitoring by eigenstructure change detection based on perturbation analysis

Participants: Michael Doehler, Qinghua Zhang, Laurent Mevel.

Vibration monitoring, notably in the fields of civil, mechanical and aeronautical engineering, aims at detecting damages at an early stage, in general by using output-only vibration measurements under ambient excitation. In this work, a new method is developed for the detection of small changes in the eigenstructure of such systems. The main idea is to transform the multiplicative eigenstructure change detection problem to an additive one, by means of perturbation analysis based on the assumption of small eigenstructure changes. Another transformation then further simplifies the detection problem into the framework of a linear regression subject to additive white Gaussian noises, leading to a numerically efficient solution of the considered problem. Compared to existing methods, it has the advantages of focusing on chosen system parameters and efficiently addressing random uncertainties. The results of this study have been presented in [31].

7.3.2. Stochastic hybrid system actuator fault diagnosis by adaptive estimation

Participant: Qinghua Zhang.

Based on the interacting multiple model (IMM) estimator for hybrid system state estimation and on the adaptive Kalman filter for time varying system joint state-parameter estimation, a new algorithm, the adaptive IMM estimator, is developed in this work for actuator fault diagnosis in stochastic hybrid systems. The working modes of the considered hybrid systems are described by stochastic state-space models, and the mode transitions are characterized by a Markov model. Actuator faults are modeled as parameter changes, and the related fault diagnosis problem is solved by the proposed adaptive IMM estimator through joint state-parameter estimation. This study has been accomplished in the framework of the ITEA MODRIO project and the results have been presented in [40].

7.3.3. Damage detection on real structures

Participants: Dominique Siegert, Laurent Mevel.

This article presents the feasibility study of a new structure for a 10-m-span bridge deck, taking into account the possibilities offered by new and high-strength materials and the advantages of a traditional environmental-friendly material. Small localized damages are hardly detected by global monitoring methods. The effectiveness of vibration-based detection depends on the accuracy of the modal parameter estimates and is limited by the low sensitivity of the modal parameters to a local stiffness reduction. This paper presents the application of SSDD to detect the change of the modal parameters of the investigated structure. Further analysis with a finite element model was conducted for assessing the consistency of the expected location and extent of the damaged elements. [15].

7.3.4. Damage detection and simulated validation

Participants: Michael Doehler, Laurent Mevel, Saeid Allahdadian.

This section is devoted to the numerical and theoretical validation of stochastic subspace damage detection. Sample length and sensor noise robustness were investigated. [24], [23], [25].

7.3.5. Damage quantification

Participants: Michael Doehler, Laurent Mevel.

Fault detection for structural health monitoring has been a topic of much research during the last decade. Localization and quantification of damages, which are linked to fault isolation, have proven to be more challenging, and at the same time of higher practical impact. While damage detection can be essentially handled as a data-driven approach, localization and quantification require a strong connection between data analysis and physical models. This paper builds upon a hypothesis test that checks if the mean of a Gaussian residual vector whose parameterization is linked to possible damage locations has become non-zero in the faulty state. It is shown how the damage location and extent can be inferred and robust numerical schemes for their estimation are derived based on QR decompositions and minmax approaches. Finally, the relevance of the approach is assessed in numerical simulations of two structures.[30].

7.3.6. Optical fiber for damage detection

Participant: Dominique Siegert.

A technique has been developed to detect and quantify structural damages. It consists of updating the model parameters associated to the damage, i.e. Young modulus, from strain sensor outputs obtained by optical fiber. Early damage detection can be expected using the local information given by the strain measurement. The method has been applied to a 8 meter post-tensioned concrete beam under a static loading. The model updating problem can be formulated as a minimization problem, i.e minimize a data misfit functional. To solve this problem, we use a gradient-based method. The gradient of the functional is computed at a low computational cost by means of the adjoint state. The technique is able to detect the damaged area in a post-tensioned concrete beam and to estimate its level of damage. [38]

7.4. Smarts roads and R5G

7.4.1. Positive surface temperature pavement

Participants: Jean Dumoulin, Nicolas Le Touz.

The mobility during winter season in France mainly relies on the use of de-icers, with an amount ranging from two hundreds thousands tons up to two millions tons for the roads only. Besides the economic impact, there are many concerns on their environmental and infrastructure, both on roads and on airports. In such context and in the framework of the R5G (5th Generation Road) project driven by IFSTTAR, investigations were carried out on the way to modify the infrastructure to maintain pavement surface at a temperature above water freezing point. Two distinct approaches, that can could be combined, were selected. The first one consisted in having a heated fluid circulating in a porous layer within an asphalt concrete pavement sample. The second one specifically relied on the use of paraffin phase change materials (PCM) in cement concrete pavement ones. Experiments on enhanced pavement samples were conducted in a climatic chamber to simulate winter conditions for several continuous days, including wind and precipitations, and monitored by infrared thermography. [45], [34]

7.4.2. Road structure design with energy harvesting capabilities

Participants: Nicolas Le Touz, Jean Dumoulin.

Facing the heavy organisational, financial and environmental constraints imposed by usual winter maintenance salting operations, pavement engineers have been led to look for alternative solutions to avoid ice or snow deposit at pavements surface. Among the solutions, one is self-de-icing heating pavements, for which two technologies have been developed so far: one is based on embedded coils circulating a heated calorific fluid under the pavement surface; the other one relies on the use of embedded resistant electric wires. The use and operation of such systems in the world is still limited and was only confined to small road stretches or specific applications, such as bridges which are particularly sensitive to frost. One of the most significant coil technology example in Europe is the SERSO-System (Solar Energy recovery from road surfaces) built in 1994, on a Switzerland bridge. Many of these experiences are referenced in the technical literature, which provides state-of-the art papers (see for instance Eugster) and useful detailed information dealing with the construction and operational management of such installation. The present study is taking part of the Forever Open Road Concept addressed by the R5G: 5th Generation Road , one of the major project supported by IFSTTAR. It considers a different design of self-de-icing road that simplify its mode of construction and maintenance, compared to the two technologies mentioned above. It should also be noted that similar to pavements instrumented with coils, such structure could be used in the reversible way to capture the solar energy at the pavement surface during sunny days and store it, to either warm the pavement at a later stage or for exogenous needs (e.g. contribution to domestic hot water). To complete our study we also considered the use of semi-transparent pavement course wearing in place of the traditional opaque one. In the present study, a 2D model was developed using FEM approach. It combines 2 numerical models. One is dedicated to the calculation of the heat transfer inside the porous layer between the fluid and the structure according to the geometry studied and the physical properties of the components of the system. The second one addresses the heat transfer inside the different layer of the pavement and was adapted to allow the insertion of a semi-transparent surface layer (for sun radiation). The temperature spatial distribution within the structure and its surface is calculated at different time step according to the evolution of boundary conditions at its surface. Various location in France were selected and calculation of the temperature field was carried-out over a year. Discussion on the performances of such system versus its location is proposed. Influence of a semi-transparent layer is also discussed. Future works will compared numerical simulations with experiments thank to a dedicated test bench under development and that will allow to test various structure in parallel. [32]

7.5. Non Destructive Testing using Infrared Thermography

7.5.1. Optimal designs of experience for thermal NDT

Participants: Antoine Crinière, Jean Dumoulin.

During previous works, square pulsed thermography was used to carry out non destructive testing of bonding quality of CFRP glued on civil engineering structures during reinforcement operations. The use of such wave form excitation was motivated by “on-site” requirements, but also by measurements duration, number of composite layers to test, depth of possible faulting areas versus temperature elevation allowed at composite level according to inner heat diffusion. Nevertheless, square pulsed excitation implies to choose an adapted heat duration. This duration is directly linked to the reliability of the parameter estimator. According to these observations, an indicator able to predict the sufficient heating time when the reliability of the parameter estimator reached an asymptotic evolution behaviour was studied. Based on the absolute thermal contrast, the proposed indicator Iph is defined with the maximum thermal contrast and the time delay between the heating time and the appearance of the maximum contrast. This indicator allows to take into account the detectability as well as the induced flaw temporal effect on the thermal contrast shape evolution. This paper will present the establishment of this indicator for optimal square heating time and present an analysis of results obtained with numerical simulations and laboratory experiments. [28]

7.5.2. Thermal NDT and signal processing

Participant: Jean Dumoulin.

This work deals with the detection of non-emergent small structures like mosaic, hidden under a plaster layer, with various spatial layout and nature. Three post processing approach by PPT, SVD and Polynomial analysis were conducted on this experimental and simulated data set. Results obtained are analysed and discussed. Finally, influence of IR camera used will be also addressed and discussed in the dissertation. [35]

7.6. Outdoor InfraRed Thermography

7.6.1. Vision enhancement through Infrared imaging for transport infrastructures

Participant: Jean Dumoulin.

Fog conditions are the cause of severe car accidents in European western countries because of the poor induced visibility. Its occurrence and intensity are still very difficult to forecast for weather services. Furthermore, visibility determination relies on expensive instruments and does not ease their dissemination. Lately, it has been demonstrated the benefit of infrared cameras to detect and to identify objects in fog while visibility is too low for eye detection. Over the past years, such cameras have become more cost effective. A research program between IFSTTAR and Cerema studied the possibility to retrieve visibility distance in a fog tunnel during its natural dissipation. The purpose of this work is to retrieve atmospheric visibility with a technique based on the combined use of infrared thermography, Principal Components Analysis (PCA) and Partial Least-Square (PLS) regression applied to infrared images.[44] and [17]

7.6.2. Outdoor thermal monitoring of large scale structures by infrared thermography

Participants: Jean Dumoulin, Antoine Crinière.

With the constant increase of the road traffic coupled with the ageing of transport infrastructure, studying and developing robust system which allows to monitor and assess those structures is of growing interest. Among the techniques used [1], thermal monitoring with infrared thermography appears to be a good compromise between a non-intrusive method and possible added value after post-processing of acquired data. Through the past decade studies have shown the ability to monitor concrete and asphalt structure by active IR thermography. On site measurement using passive thermography have also been studied, by applying qualitative methods and quantitative one. These methods have been used to perform punctual control of various duration (few hours to few days). However, infrared thermography, when it is used in a quantitative mode (not in laboratory conditions) and not in a qualitative mode (vision applied to survey), needs to process thermal radiative corrections on the raw data acquired in real time, to take into account the influences of the natural environment's evolution with time. The ICT system called "IrLaW" is based on a multi sensing approach. It connects and synchronizes information acquired by a weather station, a GPS and an infrared camera. To fulfill ICT objectives (OGCcompliant), a specific hardware architecture was also designed and studied to allow the whole system integration in a TCP/IP network. [29]

IPSO Project-Team

5. New Results

5.1. Uniformly accurate numerical schemes for highly-oscillatory Klein-Gordon and nonlinear Schrödinger equation

The work [20] 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. 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.

5.2. Higher-order averaging, formal series and numerical integration III: error bounds

In earlier works, 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 [21], 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.

5.3. Stroboscopic averaging for the nonlinear Schrödinger equation

In [18], we are concerned with an averaging procedure, -namely Stroboscopic averaging-, for highly-oscillatory evolution equations posed in a (possibly infinite dimensional) Banach space, typically partial differential equations (PDEs) in a high-frequency regime where only one frequency is present. We construct a high-order averaged system whose solution remains exponentially close to the exact one over long time intervals, possesses the same geometric properties (structure, invariants,...) as compared to the original system, and is non-oscillatory. We then apply our results to the nonlinear Schrdinger equation on the d-dimensional torus T^d, or in R^d with a harmonic oscillator, for which we obtain a hierarchy of Hamiltonian averaged models. Our results are illustrated numerically on several examples borrowed from the recent literature.

5.4. Uniformly accurate time-splitting schemes for NLS in the semiclassical limit

In [42], we construct new numerical methods for the nonlinear Schrödinger equation in the semiclassical limit. We introduce time-splitting schemes for a phase-amplitude reformulation of the equation where the dimensionless Planck constant is not a singular parameter anymore. Our methods have an accuracy which is spectral in space, of second or fourth-order in time, and independent of the Planck constant before the formation of caustics. The scheme of second-order preserves exactly the L^2 norm of the solution, as the flow of the nonlinear Schrödinger equation does. In passing, we introduce a new time-splitting method for the eikonal equation, whose precision is spectral in space and of second or fourth-order in time.

5.5. Gyroaverage operator for a polar mesh

In [33], we are concerned with numerical approximation of the gyroaverage operators arising in plasma physics to take into account the effects of the finite Larmor radius corrections. This work extended a previous approach to polar geometries. A direct method is proposed in the space configuration which consists in integrating on the gyrocircles using interpolation operator (Hermite or cubic splines). Numerical comparisons with a standard method based on a Padé approximation are performed: (i) with analytical solutions, (ii) considering the 4D drift-kinetic model with one Larmor radius and (iii) on the classical linear DIII-D benchmark case. In particular, we show that in the context of a drift-kinetic simulation, the proposed method has similar computational cost as the standard method and its precision is independent of the radius.

5.6. Asymptotic Preserving scheme for a kinetic model describing incompressible fluids

The kinetic theory of fluid turbulence modeling developed by Degond and Lemou "Turbulence models for incompressible fluids derived from kinetic theory" (J. Math. Fluid Mech. 2002) is considered for further study, analysis and simulation. Starting with the Boltzmann like equation representation for turbulence modeling, a relaxation type collision term is introduced for isotropic turbulence. In order to describe some important turbulence phenomenology, the relaxation time incorporates a dependency on the turbulent microscopic energy and this makes difficult the construction of efficient numerical methods. To investigate this problem, we focus here on a multi-dimensional prototype model and first propose an appropriate change of frame that makes the numerical study simpler. Then, a numerical strategy to tackle the stiff relaxation source term is introduced in the spirit of Asymptotic Preserving Schemes. Numerical tests are performed in a one-dimensional framework on the basis of the developed strategy to confirm its efficiency.

5.7. Numerical schemes for kinetic equations in the diffusion and anomalous diffusion limits. Part I: the case of heavy-tailed equilibrium

In [44], we propose some numerical schemes for linear kinetic equations in the diffusion and anomalous diffusion limit. When the equilibrium distribution function is a Maxwellian distribution, it is well known that for an appropriate time scale, the small mean free path limit gives rise to a diffusion type equation. However, when a heavy-tailed distribution is considered, another time scale is required and the small mean free path limit leads to a fractional anomalous diffusion equation. Our aim is to develop numerical schemes for the original kinetic model which works for the different regimes, without being restricted by stability conditions of standard explicit time integrators. First, we propose some numerical schemes for the diffusion asymptotics; then, their extension to the anomalous diffusion limit is studied. In this case, it is crucial to capture the effect of the large velocities of the heavy-tailed equilibrium, so that some important transformations of the schemes derived for the diffusion asymptotics are needed. As a result, we obtain numerical schemes which enjoy the Asymptotic Preserving property in the anomalous diffusion limit, that is: they do not suffer from the restriction on the time step and they degenerate towards the fractional diffusion limit when the mean free path goes to zero. We also numerically investigate the uniform accuracy and construct a class of numerical schemes satisfying this property. Finally, the efficiency of the different numerical schemes is shown through numerical experiments.

5.8. Comparison of numerical solvers for anisotropic diffusion equations arising in plasma physics

In [25], we are concentrated to the comparison of numerical schemes to approximate anisotropic diffusion problems arising in tokamak plasma physics. We focus on the spatial approximation by using finite volume method and on the time discretization. This latter point is delicate since the use of explicit integrators leads to a severe restriction on the time step. Then, implicit and semi-implicit schemes are coupled to finite volumes space discretization and are compared for some classical problems relevant for magnetically confined plasmas.

It appears that the semi-implicit approaches (using ARK methods or directional splitting) turn out to be the most efficient on the numerical results, especially when nonlinear problems are studied on refined meshes, using high order methods in space.

5.9. Hamiltonian splitting for the Vlasov-Maxwell equations

In [23], a new splitting is proposed for solving the Vlasov-Maxwell system. This splitting is based on a decomposition of the Hamiltonian of the Vlasov–Maxwell system and allows for the construction of arbitrary high order methods by composition (independent of the specific deterministic method used for the discretization of the phase space). Moreover, we show that for a spectral method in space this scheme satisfies Poisson's equation without explicitly solving it. Finally, we present some examples in the context of the time evolution of an electromagnetic plasma instability which emphasizes the excellent behavior of the new splitting compared to methods from the literature.

5.10. Multiscale numerical schemes for kinetic equations in the anomalous diffusion limit

In [24], we construct numerical schemes to solve kinetic equations with anomalous diffusion scaling. When the equilibrium is heavy-tailed or when the collision frequency degenerates for small velocities, an appropriate scaling should be made and the limit model is the so-called anomalous or fractional diffusion model. Our first scheme is based on a suitable micro-macro decomposition of the distribution function whereas our second scheme relies on a Duhamel formulation of the kinetic equation. Both are Asymptotic Preserving (AP): they are consistent with the kinetic equation for all fixed value of the scaling parameter $\varepsilon > 0$ and degenerate into a consistent scheme solving the asymptotic model when epsilon tends to 0. The second scheme enjoys the stronger property of being uniformly accurate (UA) with respect to epsilon. The usual AP schemes known for the classical diffusion limit cannot be directly applied to the context of anomalous diffusion scaling, since they are not able to capture the important effects of large and small velocities. We present numerical tests to highlight the efficiency of our schemes.

5.11. High-order Hamiltonian splitting for Vlasov-Poisson equations

In [40], we consider the Vlasov-Poisson equation in a Hamiltonian framework and derive new time splitting methods based on the decomposition of the Hamiltonian functional between the kinetic and electric energy. Assuming smoothness of the solutions, we study the order conditions of such methods. It appears that these conditions are of Runge-Kutta-Nyström type. In the one dimensional case, the order conditions can be further simplified, and efficient methods of order 6 with a reduced number of stages can be constructed. In the general case, high-order methods can also be constructed using explicit computations of commutators. Numerical results are performed and show the benefit of using high-order splitting schemes in that context. Complete and self-contained proofs of convergence results and rigorous error estimates are also given.

5.12. Asymptotic Preserving numerical schemes for multiscale parabolic problems

In [45], we consider a class of multiscale parabolic problems with diffusion coefficients oscillating in space at a possibly small scale ε. Numerical homogenization methods are popular for such problems, because they capture efficiently the asymptotic behaviour as $\varepsilon \to 0$, without using a dramatically fine spatial discretization at the scale of the fast oscillations. However, known such homogenization schemes are in general not accurate for both the highly oscillatory regime $\varepsilon \to 0$ and the non oscillatory regime $\varepsilon \to 1$. In this paper, we introduce an Asymptotic Preserving method based on an exact micro-macro decomposition of the solution which remains consistent for both regimes.

5.13. Parallelization of an advection-diffusion problem arising in edge plasma physics using hybrid MPI/OpenMP programming

In [35], we present a hybrid MPI/OpenMP parallelization strategy for an advection-diffusion problem, arising in a scientific application simulating tokamak's edge plasma physics. This problem is the hotspot of the system of equations numerically solved by the application. As this part of the code is memory-bandwidth limited, we show the benefit of a parallel approach that increases the aggregated memory bandwidth in using multiple computing nodes. In addition, we designed some algorithms to limit the additional cost, induced by the needed extra inter nodal communications. The proposed solution allows to achieve good scalings on several nodes and to observe 70% of relative efficiency on 512 cores. Also, the hybrid parallelization allows to consider larger domain sizes, unreachable on a single computing node.

5.14. Numerical schemes for kinetic equations in the anomalous diffusion limit. Part II: degenerate collision frequency

In [43], which is the continuation of [44], we propose numerical schemes for linear kinetic equation which are able to deal with the fractional diffusion limit. When the collision frequency degenerates for small velocities it is known that for an appropriate time scale, the small mean free path limit leads to an anomalous diffusion equation. From a numerical point of view, this degeneracy gives rise to an additional stiffness that must be treated in a suitable way to avoid a prohibitive computational cost. Our aim is therefore to construct a class of numerical schemes which are able to undertake these stiffness. This means that the numerical schemes are able to capture the effect of small velocities in the small mean free path limit with a fixed set of numerical parameters. Various numerical tests are performed to illustrate the efficiency of our methods in this context.

5.15. Analysis of the Monte-Carlo error in a hybrid semi-Lagrangian scheme

In [17] we consider Monte-Carlo discretizations of partial differential equations based on a combination of semi-lagrangian schemes and probabilistic representations of the solutions. The goal of this paper is twofold. First we give rigorous convergence estimates for our algorithm: In a simple setting, we show that under an anti-CFL condition on the time-step δt and on the mesh size δx and for a reasonably large number of independent realizations N, we control the Monte-Carlo error by a term of order $\mathcal{O}(\sqrt{\delta t/N})$. Then, we show various applications of the numerical method in very general situations (nonlinear, different boundary conditions, higher dimension) and numerical examples showing that the theoretical bound obtained in the simple case seems to persist in more complex situations.

5.16. Resonant time steps and instabilities in the numerical integration of Schrödinger equations

In [30], we consider the linear and non linear cubic Schrödinger equations with periodic boundary conditions, and their approximations by splitting methods. We prove that for a dense set of arbitrary small time steps, there exists numerical solutions leading to strong numerical instabilities preventing the energy conservation and regularity bounds obtained for the exact solution. We analyze rigorously these instabilities in the semi-discrete and fully discrete cases.

5.17. Collisions of almost parallel vortex filaments

In [38], we investigate the occurrence of collisions in the evolution of vortex filaments through a system introduced by Klein, Majda and Damodaran, and by Zakharov. We first establish rigorously the existence of a pair of almost parallel vortex filaments, with opposite circulation, colliding at some point in finite time. The collision mechanism is based on the one of the self-similar solutions of the model, described in our previous work. In the second part of this paper we extend this construction to the case of an arbitrary number of filaments, with polygonal symmetry, that are perturbations of a configuration of parallel vortex filaments forming a polygon, with or without its center, rotating with constant angular velocity.

5.18. On numerical Landau damping for splitting methods applied to the Vlasov-HMF model

In [49] we consider time discretizations of the Vlasov-HMF (Hamiltonian Mean-Field) equation based on splitting methods between the linear and non-linear parts. We consider solutions starting in a small Sobolev neighborhood of a spatially homogeneous state satisfying a linearized stability criterion (Penrose criterion). We prove that the numerical solutions exhibit a scattering behavior to a modified state, which implies a nonlinear Landau damping effect with polynomial rate of damping. Moreover, we prove that the modified state is close to the continuous one and provide error estimates with respect to the time stepsize.

5.19. A kinetic model for the transport of electrons in a graphen layer

In [50], a kinetic model for the transport of electrons in graphene is derived with the tools of semiclassical analysis. The underlying quantum model is a massless Dirac equation, whose eigenvalues display a conical singularity responsible for non adiabatic transitions between the two modes. Our kinetic model takes the form of two Boltzmann equations coupled by a collision operator modeling these transitions. This collision term includes a Landau-Zener transfer term and a jump operator whose presence is essential in order to ensure a good energy conservation during the transitions. We propose an algorithmic realization of the semi-group solving the kinetic model, by a particle method. In the last section, a series of numerical experiments are given in order to study the influences of the various sources of errors between the quantum and the kinetic models.

5.20. Dimension reduction for dipolar Bose-Einstein condensates in the strong interaction regime

In [39], we study dimension reduction for the three-dimensional Gross-Pitaevskii equation with a long-range and anisotropic dipole-dipole interaction modeling dipolar Bose-Einstein condensation in a strong interaction regime. The cases of disk shaped condensates (confinement from dimension three to dimension two) and cigar shaped condensates (confinement to dimension one) are analyzed. In both cases, the analysis combines averaging tools and semiclassical techniques. Asymptotic models are derived, with rates of convergence in terms of two small dimensionless parameters characterizing the strength of the confinement and the strength of the interaction between atoms.

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

The "interaction picture" (IP) method studied in [13] is a very promising alternative to Split-Step methods for solving certain type of partial differential equations such as the nonlinear Schrödinger equation used 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 this work 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.

5.22. Nonlinear stability criteria for the HMF Model

In [52], we study the nonlinear stability of a large class of inhomogeneous steady state solutions to the Hamiltonian Mean Field (HMF) model. Under a simple criterion, we prove the nonlinear stability of steady states which are decreasing functions of the microscopic energy. To achieve this task, we extend to this context the strategy based on generalized rearrangement techniques which was developed recently for the gravitational Vlasov-Poisson equation. Explicit stability inequalities are established and our analysis is able to treat non compactly supported steady states to HMF, which are physically relevant in this context but induces additional difficulties, compared to the Vlasov-Poisson system.

5.23. Dimension reduction for rotating Bose-Einstein condensates with anisotropic confinement

In [54], we consider the three-dimensional time-dependent Gross-Pitaevskii equation arising in the description of rotating Bose-Einstein condensates and study the corresponding scaling limit of strongly anisotropic confinement potentials. The resulting effective equations in one or two spatial dimensions, respectively, are rigorously obtained as special cases of an averaged three dimensional limit model. In the particular case where the rotation axis is not parallel to the strongly confining direction the resulting limiting model(s) include a negative, and thus, purely repulsive quadratic potential, which is not present in the original equation and which can be seen as an effective centrifugal force counteracting the confinement.

5.24. Dimension reduction for anisotropic Bose-Einstein condensates in the strong interaction regime

In [14], we study the problem of dimension reduction for the three dimensional Gross-Pitaevskii equation (GPE) describing a Bose-Einstein condensate confined in a strongly anisotropic harmonic trap. Since the gas is assumed to be in a strong interaction regime, we have to analyze two combined singular limits: a semi-classical limit in the transport direction and the strong partial confinement limit in the transversal direction. We prove that both limits commute together and we provide convergence rates. The by-products of this work are approximated models in reduced dimension for the GPE, with a priori estimates of the approximation errors.

5.25. Models of dark matter halos based on statistical mechanics: I. The classical King model

In [22], we consider the possibility that dark matter halos are described by the Fermi-Dirac distribution at finite temperature. This is the case if dark matter is a self-gravitating quantum gas made of massive neutrinos at statistical equilibrium. This is also the case if dark matter can be treated as a self-gravitating collisionless gas experiencing Lynden-Bell's type of violent relaxation. In order to avoid the infinite mass problem and carry out a rigorous stability analysis, we consider the fermionic King model. In this paper, we study the non-degenerate limit leading to the classical King model. This model was initially introduced to describe globular clusters. We propose to apply it also to large dark matter halos where quantum effects are negligible. We determine the caloric curve and study the thermodynamical stability of the different configurations. Equilibrium states exist only above a critical energy E_c in the microcanonical ensemble and only above a critical temperature T_c in the canonical ensemble.

5.26. Numerical study of a quantum-diffusive spin model for two-dimensional electron gases

In [15], we investigate the time evolution of spin densities in a two-dimensional electron gas subjected to Rashba spin-orbit coupling on the basis of the quantum drift-diffusive model. This model assumes the electrons to be in a quantum equilibrium state in the form of a Maxwellian operator. The resulting quantum drift-diffusion equations for spin-up and spin-down densities are coupled in a non-local manner via two spin chemical potentials (Lagrange multipliers) and via off-diagonal elements of the equilibrium spin density and spin current matrices, respectively. We present two space-time discretizations of the model, one semi-implicit and one explicit, which comprise also the Poisson equation in order to account for electron-electron interactions. In a first step pure time discretization is applied in order to prove the well-posedness of the two schemes, both of which are based on a functional formalism to treat the non-local relations between spin densities. We then use the fully space-time discrete schemes to simulate the time evolution of a Rashba electron gas confined in a bounded domain and subjected to spin-dependent external potentials. Finite difference approximations are first order in time and second order in space. The discrete functionals introduced are minimized with the help of a conjugate gradient-based algorithm, where the Newton method is applied in

order to find the respective line minima. The numerical convergence in the long-time limit of a Gaussian initial condition towards the solution of the corresponding stationary Schrödinger- Poisson problem is demonstrated for different values of the numerical parameters. Moreover, the performances of the semi-implicit and the explicit scheme are compared.

5.27. Numerical analysis of the nonlinear Schrödinger equation with white noise dispersion

In [16], we focus to the numerical study of a nonlinear Schrödinger equation in which the coefficient in front of the group velocity dispersion is multiplied by a real valued Gaussian white noise. We first perform the numerical analysis of a semi-discrete Crank-Nicolson scheme in the case when the continuous equation possesses a unique global solution. We prove that the strong order of convergence in probability is equal to one in this case. In a second step, we numerically investigate, in space dimension one, the behavior of the solutions of the equation for different power nonlinearities, corresponding to subcritical, critical or supercritical nonlinearities in the deterministic case. Numerical evidence of a change in the critical power due to the presence of the noise is pointed out.

5.28. A regularity result for quasilinear stochastic partial differential equations of parabolic type

In [27], we consider a quasilinear parabolic stochastic partial differential equation driven by a multiplicative noise and study regularity properties of its weak solution satisfying classical a priori estimates. In particular, we determine conditions on coefficients and initial data under which the weak solution is Hölder continuous in time and possesses spatial regularity that is only limited by the regularity of the given data. Our proof is based on an efficient method of increasing regularity: the solution is rewritten as the sum of two processes, one solves a linear parabolic SPDE with the same noise term as the original model problem whereas the other solves a linear parabolic PDE with random coefficients. This way, the required regularity can be achieved by repeatedly making use of known techniques for stochastic convolutions and deterministic PDEs.

5.29. Diffusion limit for the radiative transfer equation perturbed by a Wiener process

The aim of [28] is the rigorous derivation of a stochastic non-linear diffusion equation from a radiative transfer equation perturbed with a random noise. The proof of the convergence relies on a formal Hilbert expansion and the estimation of the remainder. The Hilbert expansion has to be done up to order 3 to overcome some difficulties caused by the random noise.

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

In [29], under an hypothesis of non-degeneracy of the flux, we 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 in 2003.

5.31. An integral inequality for the invariant measure of a stochastic reaction-diffusion equation

In [46], we consider a reaction-diffusion equation perturbed by noise (not necessarily white). We prove an integral inequality for the invariant measure ν of a stochastic reaction-diffusion equation. Then we discuss some consequences as an integration by parts formula which extends to ν a basic identity of the Malliavin Calculus. Finally, we prove the existence of a surface measure for a ball and a half-space of H.

5.32. Estimate for P_tD for the stochastic Burgers equation

In [47], we consider the Burgers equation on $H = L^2(0,1)$ perturbed by white noise and the corresponding transition semigroup P_t. We prove a new formula for P_tD_φ (where $\varphi : H \to \mathbb{R}$ is bounded and Borel) which depends on φ but not on its derivative. Then we deduce some new consequences for the invariant measure ν of P_t as its Fomin differentiability and an integration by parts formula which generalises the classical one for gaussian measures.

5.33. Existence of the Fomin derivative of the invariant measure of a stochastic reaction-diffusion equation

In [48], we consider a reaction-diffusion equation perturbed by noise (not necessarily white). We prove existence of the Fomin derivative of the corresponding transition semigroup P_t. The main tool is a new estimate for $P_tD\varphi$ in terms of $\|\varphi\|_{L^2(H,\nu)}$, where ν is the invariant measure of P_t.

5.34. Global behavior of N competing species with strong diffusion: diffusion leads to exclusion

In [19], we study the following problem. For a large class of models involving several species competing for a single resource in a *homogeneous* environment, it is known that the competitive exclusion principle holds: only one species survives eventually. Various works indicate though that coexistence of many species is possible when the competition occurs in a *heterogeneous* environment. We propose here a spatially heterogeneous system modeling several species competing for a single resource, and migrating in the spatial domain. For this model, it is known, at least in particular cases, that if migrations are *slow* enough, then coexistence occurs. In this paper we show at variance that if the spatial migrations are *fast* enough, then our system can be approximated by a spatially homogeneous system, called aggregated model, which can be explicitly computed, and we show that if the competitive exclusion principle holds for the aggregated model, then it holds as well for the original, spatially heterogeneous model. In other words, we show the persistence of the competitive exclusion principle in the spatially heterogeneous situation when migrations are fast. As a consequence, for fast migrations only one species may survive, namely the best competitor *in average*. We last study which is the best competitor *in average* on some examples, and draw some ecological consequences.

5.35. Randomized message-passing test-and-set

In [37] and [34], we present a solution to the well-known Test&Set operation in an asynchronous system prone to process crashes. Test&Set is a synchronization operation that, when invoked by a set of processes, returns yes to a unique process and returns no to all the others. Recently many advances in implementing Test&Set objects have been achieved, however all of them target the shared memory model. In this paper we propose an implementation of a Test&Set object in the message passing model. This implementation can be invoked by any number $p \leq n$ of processes where n is the total number of processes in the system. It has an expected individual step complexity in $O(\log p)$ against an oblivious adversary, and an expected individual message complexity in $O(n)$. The proposed Test&Set object is built atop a new basic building block, called selector, that allows to select a winning group among two groups of processes. We propose a message-passing implementation of the selector whose step complexity is constant. We are not aware of any other implementation of the Test&Set operation in the message passing model.

DYLISS Project-Team

7. New Results

7.1. Data integration

Participants: Jacques Nicolas, Charles Bettembourg, Jérémie Bourdon, Jeanne Got, Marie Chevallier, Guillaume Collet, Olivier Dameron, Damien Eveillard, Julie Laniau, Anne Siegel.

Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. Interaction graphs provide a suitable representation of cellular networks with information flows. Methods based on sign consistency have been shown to be valuable tools to (i) predict qualitative responses, (ii) test the consistency of network topologies and experimental data, and (iii) apply repair operations to the network model suggesting missing or wrong interactions. We present a framework to unify different notions of sign consistency and propose a refined method for data discretization that considers uncertainties in experimental profiles. We furthermore introduce a new constraint to filter undesired model behaviors induced by positive feedback loops. Finally, we generalize the way predictions can be made by the sign consistency approach. This corresponds to an extension of our *Bioquali* software. [*Anne Siegel*] [21]

Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. Our software tool *shogen* was used to decipher functional roles within a consortium of five mining bacteria through the integration of genomic and metabolic knowledge at genome scale. We first reconstructed a global metabolic network. Next, using a parsimony assumption, we deciphered sets of genes, called Sets from Genome Segments (SGS), that (i) are close on their respective genomes, (ii) take an active part in metabolic pathways and (iii) whose associated metabolic reactions are also closely connected within metabolic networks. The use of SGS (shogen) pinpoints a functional compartmentalization among the investigated species and exhibits putative bacterial interactions necessary for promoting these pathways. [*Damien Eveillard, Anne Siegel*] [17]

Optimal Threshold Determination for Interpreting Semantic Similarity and Particularity: Application to the Comparison of Gene Sets and Metabolic Pathways Using GO and ChEBI. We developed a method for determining optimal semantic similarity and particularity thresholds in order to interprete the results of the comparison of ontology terms sets. We applied this method on the GO and ChEBI ontologies. Qualitative analysis using the thresholds on the PPAR multigene family yielded biologically-relevant patterns. [*Charles Bettembourg, Olivier Dameron*] [16]

AskOmics : Integration et interrogation de reseaux de regulation genomique et post-genomique. We present AskOmics, an integration and interrogation software using a RDF model and the SPARQL query language. The purpose of this work is to obtain quick answers to biological questions demanding currently hours of manual search in several spreadsheet results files. AskOmics allows biologists to integrate and interrogate their data by themselves without any knowledge about RDF and SPARQL required. [*Charles Bettembourg, Olivier Dameron*] [30]

7.2. Time-series and asymptotic dynamics

Participants: Anne Siegel, Jacques Nicolas, Jérémie Bourdon, Jean Coquet, Victorien Delannée, Vincent Picard, Nathalie Théret.

Identification of logical models for signaling pathways: towards a systems biology loop. Logical models of signaling pathways are a promising way of building effective in silico functional models of a cell. The automated learning of Boolean logic models describing signaling pathways can be achieved by training to phosphoproteomics data. This data is unavoidably subject to noise. As a result, the learning process leads to a family of feasible logical networks rather than a single model. This family is composed of logic models proposing different internal wirings for the system, implying that the logical predictions from this family may suffer a significant level of variability leading to uncertainty. In our work, combinatorial optimization methods based on recent logic programming paradigm allow to enumerate, and discriminate the family of logical models explaining data. Together, these approaches enable a robust understanding of the system response. The results are implemented in the *caspo* software [*Jacques Nicolas, Anne Siegel*] [22], [23]

Boolean Network Identification from Multiplex Time Series Data. The ASP-based learning algorithm developed in the team to train logical models of signaling networks focuses on the comparison of two time-points and assumes that the system has reached an early steady state. We have generalized such a learning procedure in order to discriminate Boolean networks according to their transient dynamics. To that goal, we exhibit a necessary condition that must be satisfied by a Boolean network dynamics to be consistent with a discretized time series trace. This approach was included in the ASP-based framework designed for the caspo software. We ended up with a global learning algorithm and compared it to learning approaches based on static data. [*Anne Siegel*] [31]

Representation of symbolic dynamical systems generated by a substitution. Iterated morphisms are combinatorial processes which are related to several classes of dynamical systems appearing in several fields of computer sciences and mathematics: numeration, ergodic theory, discrete geometry. They may be associated to fractal sets called "Rauzy fractals" whose topological properties are linked to the properties of the underlying dynamical system. We have introduced a generic algorithm framework to check such topological properties within a complete family of iterated morphism. This makes efficient the verification of conjectures on several families of substitutions related to multi-dimensional continued fraction algorithms. [*Anne Siegel*] [32], [25], [14]

Multivariate Normal Approximation for the Stochastic Simulation Algorithm: Limit Theorem and Applications. We present a central limit theorem for the Gillespie stochastic trajectories when the living system has reached a steady-state, that is when the internal bio-molecules concentrations are assumed to be at equilibrium. It appears that the stochastic behavior in steady-state is entirely characterized by the stoichiometry matrix of the system and a single vector of reaction probabilities. We propose several applications of this result such as deriving multivariate confidence regions for the time course of the system and a constraints-based approach which extends the flux balance analysis framework to the stochastic case. [*Jérémie Bourdon, Vincent Picard, Anne Siegel*] [20], [12]

A Logic for Checking the Probabilistic Steady-State Properties of Reaction Networks. Designing probabilistic reaction models and determining their stochastic kinetic parameters are major issues in systems biology. In order to assist in the construction of reaction network models, we introduce a logic that allows one to express asymptotic properties about the steady-state stochastic dynamics of a reaction network. Basically, the formulas can express properties on expectancies, variances and co-variances. We demonstrate that deciding the satisfiability of a formula is NP-hard. [*Jérémie Bourdon, Vincent Picard, Anne Siegel*] [28], [12]

7.3. Sequence and structure annotation

Participants: François Coste, Aymeric Antoine-Lorquin, Catherine Belleannée, Guillaume Collet, Clovis Galiez, Laurent Miclet, Jacques Nicolas.

Amplitude Spectrum Distance: measuring the global shape divergence of protein fragments. We introduce here the Amplitude Spectrum Distance (ASD), a novel way of comparing protein fragments based on the discrete Fourier transform of their C_α distance matrix. Defined as the distance between their amplitude spectra, ASD can be computed efficiently and provides a parameter-free measure of the global shape dissimilarity of two fragments. ASD inherits from nice theoretical properties, making it tolerant to shifts, insertions,

deletions, circular permutations or sequence reversals while satisfying the triangle inequality. The practical interest of ASD with respect to RMSD, $RMSD_d$, BC and TM scores is illustrated through zinc finger retrieval experiments and concrete structure examples. The benefits of ASD are also illustrated by two additional clustering experiments: domain linkers fragments and complementarity-determining regions of antibodies. [*Clovis Galiez, François Coste*] [19]

Structural conservation of remote homologues: better and further in contact fragments. We address a basic question on sequence-structure relationships in proteins: does a protein sequence depict a structure with a uniform faithfulness all along the sequence ? We investigate this question by defining contact fragments. This study suggests that sequence homologs of CF are significantly more faithful to structure than randomly chosen fragments, so that CF carry a strong sequence-structure relationship, allowing them to be used as accurate building blocks for structure prediction. [*Clovis Galiez, François Coste*] [26]

VIRALpro: a tool to identify viral capsid and tail sequences. Not only sequence data continues to outpace annotation information, but the problem is further exacerbated when organisms are underrepresented in the annotation databases. This is the case with non human-pathogenic viruses which occur frequently in metagenomic projects. Thus there is a need for tools capable of detecting and classifying viral sequences. We describe VIRALpro a new effective tool for identifying capsid and tail protein sequences, which are the cornerstones toward viral sequence annotation and viral genome classification. [*Clovis Galiez, François Coste*] [18]

Finding Optimal Discretization Orders for Molecular Distance Geometry. The Molecular Distance Geometry Problem (MDGP) is the problem of finding the possible conformations of a molecule by exploiting available information about distances between some atom pairs. Under minimal assumptions the MDGP can be discretized so that the search domain of the problem becomes a tree that can be explored by using an interval Branch & Prune (iBP) algorithm. In this context, the discretization assumptions are strongly dependent on the atomic ordering, which can also impact the computational cost of the iBP algorithm. In this work, we propose a new partial discretization order for protein backbones. This new atomic order optimizes a set of objectives that aim at improving the iBP performances. The optimization of the objectives is performed by Answer Set Programming (ASP), which allows to express the problem by a set of logical constraints. The comparison with previously proposed orders for protein backbones shows that this new discretization order makes iBP perform more efficiently. [*Jacques Nicolas*] [34]

From formal concepts to analogical complexes. Reasoning by analogy is an important component of common sense reasoning whose formalization has undergone recent improvements with the logical and algebraic study of the analogical proportion. The starting point of this study considers analogical proportions on a formal context. We introduce analogical complexes, a companion of formal concepts formed by using analogy between four subsets of objects in place of the initial binary relation. They represent subsets of objects and attributes that share a maximal analogical relation. We show that the set of all complexes can be structured in an analogical complex lattice and give explicit formulae for the computation of their infimum and supremum. [*Laurent Miclet, Jacques Nicolas*] [27]

Comparison of the targets obtained by a scoring matrix and by a regular expression. Application to the search for LXR binding sites. In bioinformatics, it is a common task to search for new instances of a pattern built from a set of reference sequences. For the simplest and most frequent cases, patterns are represented in two ways : regular expression or scoring matrix. Since both representations seem to be used indifferently in pratice, one may wonder if they have any impact on the result. This study compares hits obtained with scoring matrices or by regular expressions allowing up to two substitutions. It shows that, in our LXR study, sequences found by a scoring matrix are closer to the targeted hits than sequences found by a regular expression. [*Aymeric Antoine-Lorquin, Jacques Nicolas, Catherine Belleannée*] [29]

Finding and Characterizing Repeats in Plant Genomes. Plant genomes contain a particularly high proportion of repeated structures of various types. This chapter proposes a guided tour of available softwares that can help biologists to look for these repeats and check some hypothetical models intended to characterize their structures. Since transposable elements are a major source of repeats in plants, we have provided a whole section on this topic as well as a selection of the main existing softwares. In order to better understand how

they work and how repeats may be efficiently found in genomes, the rest of the chapter is devoted to the foundations of the search for repeats and more complex patterns. We first introduce the key concepts that are useful for understanding the current state of the art in playing with words, applied to genomic sequences. In fact, biologists need to represent more complex entities where a repeat family is built on more abstract structures, including direct or inverted small repeats, motifs, composition constraints as well as ordering and distance constraints between these elementary blocks. The last section introduces concepts and practical tools that can be used to reach this syntactic level in biological sequence analysis. [*Jacques Nicolas*] [35]

FLUMINANCE Project-Team

6. New Results

6.1. Fluid motion estimation

6.1.1. Stochastic uncertainty models for motion estimation

Participants: Etienne Mémin, Abed Malti.

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. The development of an efficient GPU based version of this estimator has been investigated through the Inria ADT project FLUMILAB

6.1.2. 3D flows reconstruction from image data

Participants: Kai Berger, Cédric Herzet, Abed Malti.

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.

This year, we keep on our investigation on the problem of efficient volume reconstruction. Our work takes place within the context of some modern optimization techniques. First, we focussed our attention on the family of proximal and splitting methods and showed that the standard techniques commonly adopted in the TomoPIV literature can be seen as particular cases of such methodologies. Recasting standard methodologies in a more general framework allowed us to propose extensions of the latter: i) we showed that the parcimony characterizing the sought volume can be accounted for without increasing the complexity of the algorithms (e.g., by including simple thresholding operations); ii) we emphasized that the speed of convergence of the standard reconstruction algorithms can be improved by using Nesterov's acceleration schemes; iii) we also proposed a totally novel way of reconstructing the volume by using the so-called "alternating direction of multipliers method" (ADMM) . This work has led to the publication of two contributions at the international conference on particle image velocimetry (PIV) in 2015.

On top of this work, we also focussed on another crucial step of the volume reconstruction problem, namely the pruning of the model. The pruning task consists in identifying some positions in the volume of interest which cannot contains any particle. Removing this position from the problem can then potentially allow for a dramatic dimensionality reduction. This year, we provide a methodological answer to this problem through the prism of the so-called "screening" techniques which have been proposed in the community of machine learning. Our work has led to the submission of one contribution to the international conference on acoustics, speech and signal processing.

6.1.3. 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 access to the solution of this problem. Among the popular methods available in the literature, one can mention orthogonal matching pursuit (OMP), orthogonal least squares (OLS) 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 contributed to this research axis by deriving conditions of success for the algorithms mentioned above when the amplitudes of the nonzero coefficients in the sparse vector obey some decay. In a TomoPIV context, this decay corresponds to the fact that not all the particles in the fluid diffuse the same quantity of light (notably because of illumination or radius variation). In particular, we show that the standard coherence-based guarantees for OMP/OLS can be relaxed by an amount which depends on the decay of the nonzero coefficients. Our work have led to the acceptation of a paper in the journal IEEE Transactions on Information Theory.

6.2. Tracking, Data assimilation and model-data coupling

6.2.1. 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.

6.2.2. 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 [9]. 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. 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). This year we have more deeply explored several uncertainty representations of classical geophysical models for ocean and atmosphere. This study have led to very promising stochastic representation for the Quasi Geostophic approximation (QG) with noises of different energy.

6.2.3. Free surface flows reconstruction and tracking

Participants: Dominique Heitz, Etienne Mémin.

We investigated the combined use of a Kinect depth sensor and of a stochastic data assimilation method to recover free-surface flows. More generally, we proposed a particle filter method to reconstruct the complete state of free-surface flows from a sequence of depth images only. The data assimilation scheme introduced accounts for model and observations errors. We evaluated the developed approach on two numerical test cases: a collapse of a water column as a toy-example and a flow in an suddenly expanding flume as a more realistic flow. The robustness of the method to simulated depth data quality and also to initial conditions was considered. We illustrated the interest of using two observations instead of one observation into the correction step. Then, the performance of the Kinect sensor to capture temporal sequences of depth observations was investigated. Finally, the efficiency of the algorithm was qualified for a wave in a real rectangular flat bottom tank. It was shown that for basic initial conditions, the particle filter rapidly and remarkably reconstructed velocity and height of the free surface flow based on noisy measures of the elevation

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

Participants: Pranav Chandramouli, Dominique Heitz, Etienne Mémin, Cordelia Robinson.

In this axis of work we are exploring 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 [6]. 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 has been 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. The 4DVar assimilation technique based on the numerical code Incompact3D is now implemented. We are currently trying to reconstruct a 3D turbulent flow from dual

plane velocity observations. The control of subgrid parameterizations will be the main objective of the PhD of Pranav Chandramouli that is just starting.

6.2.5. Ensemble variational data assimilation of large scale fluid flow dynamics with uncertainty

Participant: Etienne Mémin.

This study is focused on the coupling of a large scale representation of the flow dynamics built from the location uncertainty principle with image data of finer resolution. The velocity field at large scales is described as a regular smooth component 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 assessed the performance of an ensemble variational assimilation technique with direct image data observation. Preliminary encouraging results have been obtained for simulation under uncertainty of 1D and 2D shallow water models.

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

Participants: Cédric Herzet, Etienne Mémin, Valentin Resseguier.

During the PhD thesis of Valentin Ressguier we proposed a new decomposition of the fluid velocity in terms of a large-scale continuous component with respect to time and a small-scale non continuous random component. Within this general framework, an uncertainty based representation of the Reynolds transport theorem and Navier-Stokes equations can be derived, based on physical conservation laws. This physically relevant stochastic model has been applied in the context of the POD-Galerkin method. The pertinence of this reduced order model has been successfully assessed on several wake flows. This study has been published in two conference papers and one journal article.

On the other hand, we also investigated the problem of reduced-model construction from partial observations. In this line of search, our contribution was twofold. We first proposed a Bayesian framework for the construction of reduced-order models from image data. Our framework enables to account for any prior information on the system to reduce and takes the uncertainties on the parameters of the model into account. Interestingly, the proposed approach reduces to some well-known model-reduction techniques when the observations are not partial (i.e., the observation operator can be inverted). Second, we provided a theoretical analysis of our methodology in a simplified context (namely, the observations are supposed to be noiseless linear combinations of the state of the system). This result provides worst-case guarantees on the reconstruction performance which can be achieved by a reduced model built from the data. These contributions have been accepted for presentation in two international conferences in 2016.

6.3. Analysis and modeling of turbulent flows

6.3.1. Turbulence similarity theory for the modeling of Ocean Atmosphere interface

Participants: Roger Lewandowski, Etienne Mémin, Benoit Pinier.

The Ocean Atmosphere interface plays a major role in climate dynamics. This interaction takes place in a thin turbulent layer. To date no sastifying universal models for the coupling of atmospheric and oceanic models exists. In practice this coupling is realized through empirically derived interaction bulks. In this study, corresponding to the PhD thesis of Benoit Pinier, we aim at exploring similarity theory to identify universal mean profile of velocity and temperature within the mixture layer. The goal of this work consists in exhibiting eddy viscosity models within the primitive equations. We will also explore the links between those eddy viscocity models and the subgrid tensor derived from the uncertainty framework studied in the Fluminance group. In that prospect, we have started to study the impact of the introduction of a random modeling of the friction velocity on the classical wall law expression.

6.3.2. 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.

6.3.3. Numerical and experimental image and flow database

Participants: Pranav Chandramouli, Dominique Heitz.

The goal was to design a database for the evaluation of the different techniques developed in the Fluminance group. The first challenge was to enlarge a database mainly based on two-dimensional flows, with three-dimensional turbulent flows. Synthetic image sequences based on homogeneous isotropic turbulence and on circular cylinder wake have been provided. These images have been completed with time resolved Particle Image Velocimetry measurements in wake and mixing layers flows. This database provides different realistic conditions to analyse the performance of the methods: time steps between images, level of noise, Reynolds number, large-scale images. The second challenge was to carried out orthogonal dual plane time resolved stereoscopic PIV measurements in turbulent flows. The diagnostic employed two orthogonal and synchronized stereoscopic PIV measurements to provide the three velocity components in planes perpendicular and parallel to the streamwise flow direction. These temporally resolved planar slices observations will be used in 4DVar assimilation technique, integrating Direct Numerical Simulation (DNS) and Large Eddies Simulation (LES), to reconstruct three-dimensional turbulent flows. This reconstruction will be conducted within the PhD of Pranav Chandramouli. The third challenge was to carried out a time resolved tomoPIV experiments in a turbulent wake flow. These temporally resolved volumic observations will be used to assess the algorithms developped in the PhD of Ioana Barbu and in the postdoc of Kai Berger. Then this data will be used in 4DVar assimilation technique to reconstruct three-dimensional turbulent flows. This reconstruction will be conducted within the PhD of Cordelia Robinson.

6.4. Visual servoing approach for fluid flow control

6.4.1. Closed-loop control of a spatially developing shear layer

Participant: Christophe Collewet.

This study is led within a strong collaboration with Diemer Ando-Ondo and Johan Carlier of the Acta team (Irstea Rennes). It aims at controling one of the prototypical flow configurations encountered in fluid mechanics: the spatially developing turbulent shear layer occuring between two parallel incident streams with different velocities. Closed loop control is achieved to maintain the shear-layer in a desired state of interest for industrial applications, and thus to reject upstream perturbations. The industrial and scientific contexts advocates first for the use of image sensor to measure the flow velocity fields and second for applying the control on the upstream boundary condition. The optimal control was performed using a linear control law designed from a reduced linearized state space model of the Navier-Stokes equations. A steady desired state was first considered leading to a linear time-invariant system. The resulting feedback control law was validated on a powerful and realistic numerical Navier-Stokes 3D solver, which will be useful to anticipate the control of the shear layer in a dedicated wind tunnel. Two conference papers on this work have been submitted to the "16th European Control Conference" and "8th AIAA Flow Control Conference". We are now considering the case of an unsteady desired state to control the large roller vortices developing in the shear layer and that are the main contributor to entrainment and mixing processes.

GENSCALE Project-Team

7. New Results

7.1. HTS data processing

7.1.1. Genome Analysis Tool Box Optimization

Participants: C. Deltel, P. Durand, E. Drezen, D. Lavenier, C. Lemaitre, P. Peterlongo, G. Rizk

Among the GATB library, the kmer-counting procedure is one of the most useful building block to speed-up development of new NGS tools. It is the first step of many NGS tools developed with GATB : Leon, Bloocoo, MindTheGap, DiscoSnp, Simka, TakeAbreak. This procedure has been optimized to be less limited by disks I/O. It relies on the use of kmer minimizers that help quickly partition the whole set of kmers into compact subsets. The kmer-counting procedure has also been re-worked to be more versatile, it is now able to count separately many input files and allows easy parametrization of the output, from simple kmer-count to the creation of custom user-defined kmer measures. At the core of the GATB library is also the manipulation and traversal of the de Bruijn Graph. The implementation has been optimized, leading to graph traversal twice fast as before. We introduced a new type of bloom filters, that are specially optimized for the manipulation of kmers. In these bloom filters neighboring kmers in the graph are close together in the bloom filter bit array, leading to better data locality, less cache misses and better overall performance [38].

7.1.2. NGS Data Compression

Participants: G. Benoit, E. Drezen, D. Lavenier, C. Lemaitre, G. Rizk

A novel reference-free method to compress data issued from high throughput sequencing technologies has been developed. Our approach, implemented in the LEON software, employs techniques derived from assembly principles. The method is based on a reference probabilistic de-Bruijn Graph, built de novo from the set of reads and stored in a Bloom filter. Each read is encoded as a path in this graph, by memorizing an anchoring kmer and a list of bifurcations. The same probabilistic de-Bruijn Graph is used to perform a lossy transformation of the quality scores allowing higher compression rates to be obtained without loosing pertinent information for downstream analyses. Leon was run on various real sequencing datasets (whole genome, exome, RNA-seq or metagenomics). In all cases, LEON showed higher overall compression ratios than state-of-the-art compression software. On a *C. elegans* whole genome sequencing dataset, LEON divided the original file size by more than 20 [16].

7.1.3. Multistep global optimization approach for the scaffolding problem

Participants: R. Andonov, D. Lavenier, I. Petrov

Our overall goal here is to address the computational hardness of the scaffolding problem by designing faster algorithms for global optimization that combine the branch-and-bound method which is able to find the global optimum but is usually slow for accuracy, with the use of massive parallelism and exploiting the special properties of the data–for scalability. A new two step scaffolding modeling strategy is in development. It tries to break the problem complexity by first solving a graph containing only large unitigs building something that can be compared to a trustworthy genomic frame. In our preliminary works [40] we developed integer programming optimization models that have been successfully applied on synthetic data generated from small chloroplast genomes. For computation we uses the Gurobi optimization solver.

7.1.4. Mapping reads on graph

Participants: A. Limasset, C. Lemaitre, P. Peterlongo

Next Generation Sequencing (NGS) has dramatically enhanced our ability to sequence genomes, but not to assemble them. In practice, many published genome sequences remain in the state of a large set of contigs. Although many subsequent analyses can be performed, one may ask whether mapping reads on the contigs is as informative as mapping them on the paths of the assembly graph. We proposed a formal definition of mapping on a de Bruijn graph, analysed the problem complexity which turned out to be NP-complete, and provided a practical solution. We proposed a pipeline called GGMAP (Greedy Graph MAPping). Its novelty is a procedure to map reads on branching paths of the graph, for which we designed a heuristic algorithm called BGREAT (de Bruijn Graph REAd mapping Tool). For the sake of efficiency, BGREAT rewrites a read sequence as a succession of unitigs sequences. GGMAP can map millions of reads per CPU hour on a de Bruijn graph built from a large set of human genomic reads. Surprisingly, results show that up to 22% more reads can be mapped on the graph but not on the contig set. Although mapping reads on a de Bruijn graph is a complex task, our proposal offers a practical solution combining efficiency with an improved mapping capacity compared to assembly-based mapping even for complex eukaryotic data [43].

7.1.5. Improving discoSnp features

Participants: C. Riou, C. Lemaitre, P. Peterlongo

NGS data enable to detect polymorphisms such as SNPs and indels. Their detection in NGS data is now a routine task. The main methods for their prediction usually need a reference genome. However, non-model organisms and highly divergent genomes such as in cancer studies are more and more investigated. The discoSnp tool has been successfully applied to predict isolated SNPs from raw read set(s) without the need of a reference genome. We improved discoSnp which became discoSnp++ [44]. DiscoSnp++ benefits from a new software design that reduces time and memory consumption, and from a new algorithmic design that detects all kinds of SNP and small indels, adds genotype information and outputs a VCF (Variant Calling Format) file. Moreover, when a reference genome may be used, discoSnp++ predictions are automatically mapped to this reference and the VCF file shows up location information of each prediction. This step also provides a way to filter out false predictions due to genomic repeats. Using discoSnp++ even when a reference is available has multiple advantages: it is several order of magnitude faster and uses much less memory. We are currently working in showing that it also provides better predictions than methods based on read mapping.

7.1.6. HLA genotyping

Participant: D. Lavenier

The human leukocyte antigen (HLA) system drives the regulation of the Human immune system. Genotyping the HLA genes involved in the immune system consists first in a deep sequencing of the HLA region. Next, a NGS analysis is performed to detects SNP variations from which correct haplotypes are computed. We have developed a fast method that outperforms standard approaches which, generally, require exhaustive database searches. Instead, the method extracts a few significant k-mers from all the haplotypes referenced in the HLA database. Each haplotype is then characterized by a small set of informative k-mers. By comparing these k-mer sets with the HLA sequencing data of a specific person, we can rapidly determine its HLA genotype.

7.1.7. Identification of long non-coding RNAs in insects genomes

Participant: F. Legeai

The development of high throughput sequencing technologies (HTS) has allowed researchers to better assess the complexity and diversity of the transcriptome. Among the many classes of non-coding RNAs (ncRNAs) that were identified during the last decade, long non-coding RNAs (lncRNAs) represent a diverse and numerous repertoire of important ncRNAs, reinforcing the view that they are of central importance to the cell machinery in all branches of life. Although lncRNAs have been involved in essential biological processes such as imprinting, gene regulation or dosage compensation especially in mammals, the repertoire of lncRNAs is poorly characterized for many non-model organisms [23]. In collaboration with the Institut de Génétique et de Développement de Rennes (IGDR) we participate in the development of a software for extracting long non coding RNA from high throughput data (https://github.com/tderrien/FEELnc).

7.1.8. Data-mining applied to GWAS

Participants D. Lavenier, Pham Hoang Son

Discriminative pattern mining methods are powerful techniques for discovering variant combinations related to diseases. The aim is to find a set of patterns that occur with disproportion frequency in case-control data sets, and a real challenge is to select a complete set of variant combinations that are biologically significant. There are various measurement methods for evaluating the discriminative power of individual combination in two-class data sets. Our research activity on this topic attempts to compare the statistical discriminative power measurements in genetic case-control data sets in order to evaluate the effectiveness of detecting variants associated with diseases.

7.2. Sequence comparison

7.2.1. Amplicon alignment

Participants: S. Brillet, C. Deltel, P. Durand, D. Lavenier, I. Petrov

Many metagenomics projects identify species by the studying16S-RNA sequences. This is mainly done by comparing the amplicons with 16S-RNA bacterial banks (amplicons are short fragments sequenced from very specific genome areas). As these sequences share a lot of similarities, immediate blast-like heuristics achieve poor performances. To speed up the process, we first select informative k-mers, from both the amplicon dataset and in the RNA16 bank (informative k-mers are defined as under represented k-mers). An index is built from this reduced set of k-mers and a "seed-and-extend" procedure is run. This strategy avoids many non-useful computation and accelerate the overall computation by two orders of magnitude. This new approach is currently implemented in the PLAST software (Regional KoriPlast2 project).

7.2.2. Metagenomics datasets comparison

Participants: G. Benoit, D. Lavenier, C. Lemaitre, P. Peterlongo, G. Rizk

We developp a new method, called Simka, to compare simultaneously numerous large metagenomics datasets. The method computes pairwise distances based on the amount of shared k-mers between datasets. The method scales to a large number of datasets thanks to an efficient kmer-counting step that processes all datasets simultaneoulsy. Additionnally, several distance definitions were implemented and compared, including some originating from the ecological domain. The method is currently applied to the TARA oceans project (more than 500 datasets) which aims at comparing worldwide sea water samples (ANR HydrGen project) [39].

7.3. Protein 3D structure

7.3.1. Discovering protein conformations by distance geometry

Participant: A. Mucherino

The distance geometry asks whether a simple weighted undirected graph G can be embedded in a Euclidean space having a predefined dimension $K > 0$, so that distances between pairs of embedded vertices are the same as the weights on graph edges. One of the most important applications of the distance geometry can be found in biology, where experimental techniques are able to find estimates of certain distances between atom pairs in molecules. Even if the scientific community is used to employ standardized techniques for the solution of this problem, which are essentially based on heuristic searches, we have recently shown that our combinatorial approach to this problem can be in fact employed for solving biological instances of the distance geometry [17]. This work is in collaboration with international people and researchers from the Pasteur Institut in Paris.

7.3.2. Discretization orders for distance geometry

Participant: A. Mucherino

The concept of discretization order is fundamental for the discretization of the distance geometry, i.e. for reducing the search space of a given distance geometry instance to a discrete (and finite) space. A discretization order is an order on the vertices of the graph G representing an instance of the distance geometry that is able to satisfy the discretization assumptions. Recent research was focused on the problem of finding, for a given distance geometry instance, a suitable discretization order that allows for its discretization [32]. The problem is tackled from a purely theoretical point of view in [33], while a special order for protein backbones was identified in [27] by creating a path on a "pseudo" de Bruijn graph. In [36], additional requirements are included during the search for a vertex order, in order to identify discretization orders that are also "optimal". In this work, we used Answer Set Programming (ASP) for identifying optimal partial orders that ensure the discretization of distance geometry instances related to proteins. This work is in collaboration with the Dyliss team, as well as international people.

7.3.3. Structure Similarity Detection

Participants: M. Le Boudic-jamin, R. Andonov

The most commonly used among the various measures of alignment similarity are the internal distances root mean squared deviation (RMSDd) and the coordinate root mean squared deviation (RMSDc) . In the paper [18] we introduce a novel approach to find similarities between protein structures. Our algorithm is both internal-distances based and Euclidean-coordinates based (i.e., it uses a rigid transformation to optimally superimpose the two structures). Resulting alignments are guaranteed to score well for both RMSDd and RMSDc , while remaining polynomial. We also replace the goal of finding the largest clique by the one of returning several very dense “near-clique” subgraphs. This choice is strongly justified by the observation that distinct solutions to the structural alignment problem that are close to the optimum are all equally viable from the biological perspective, and hence are all equally interesting from the computation standpoint. Our tool is suitable for detecting similar domains when comparing multi-domain proteins, as well to detect structural repetitions within a single protein and between related proteins [12].

7.3.4. Automatic Classification of Protein Structure

Participants: M. Le Boudic-jamin, R. Andonov

In this paper [15] we propose a new distance measure for comparing two protein structures based on their contact map representations . We show that our novel measure, which we refer to as the maximum contact map overlap (max-CMO) metric, satisfies all properties of a metric on the space of protein representations. Having a metric in that space allows one to avoid pairwise comparisons on the entire database and, thus, to significantly accelerate exploring the protein space compared to no-metric spaces. We show on a gold standard superfamily classification benchmark set of 6759 proteins that our exact k-nearest neighbor (k-NN) scheme classifies up to 224 out of 236 queries correctly and on a larger, extended version of the benchmark with 60 850 additional structures, up to 1361 out of 1369 queries. Our k-NN classification thus provides a promising approach for the automatic classification of protein structures based on flexible contact map overlap alignments.

7.3.5. Detection of structure repeats in proteins

Participant: M. Le Boudic-jamin, R. Andonov

Almost 25% of proteins contain internal repeats, these repeats may have a major role in the protein function. Furthermore some proteins actually are the same substructure repeated many times, these proteins are solenoids. However, very few protein repeats detection programs exist today. In the paper [29] we present a simple and efficient tool for discovering protein repeats. Our tool is based on protein fragment comparison and clique detection. We show that our tool is able to detect different levels of repetitions and to successfully identify protein tiles.

7.4. Parallelism

7.4.1. Processor in Memory

Participants: C. Deltel, D. Lavenier

The concept of PIM (Processor In Memory) aims to dispatch the computer power near the data. Together with the UPMEM company, which is currently developing a DRAM enhanced with computing units, we investigate the parallelization of several bioinformatics algorithms for this new types of memory. The first results show that blast-like algorithms or mapping algorithms can highly benefit of such memory. But the core algorithms must be revisited in order to better suite the PIM architecture.

7.4.2. Alignment search tools on cloud

Participants: S. Brillet, D. Lavenier, I. Petrov

PLAST is an alternative version of Blast to target intensive sequence comparison (bank-to-bank comparison). The multicore version offers a speed from 5 to 10 compared to Blast. In 2015, we deploy PLAST in the IFB cloud infra-structure (French Bioinformatics Institute) and demonstrate that an Hadoop implementation provides a very good scalability [34].

7.4.3. Bioinformatics Workflow

Participants: D. Lavenier, F. Moorews

Bioinformatics workflows play an important role in the development of new methodologies for analyzing sequencing data. Optimizing this activity brings the questions of how workflow can be efficiently captured and how technical tasks integration can be simplified. Thus, we define an expressive graphic worfklow language, adapted to the quick capture of workflows. This graphical input is then interpreted by a workflow engine based on a new model of computation with high performances obtained by the use of multiple levels of parallelism. A Model-Driven design approach is associated to facilitate the data parallelism generation and the production of suitable implementations for different execution contexts. In the case of the cloud model Container as a Service (CaaS), a workflow specification intrinsically re-executable and readily disseminatable has been developed. The adoption of this kind of model could lead to an acceleration of exchanges and a better availability of data analysis workflows [25] [31] [13].

7.4.4. Graph processing

Participants: D. Lavenier, R. Andonov

In the paper [20] we present a new approach for solving the all-pairs shortest-path (APSP) problem for planar graphs that exploits the massive on-chip parallelism available in today's Graphics Processing Units (GPUs). We describe two new algorithms based on our approach. Both algorithms use Floyd-Warshall method, have near optimal complexity in terms of the total number of operations, while their matrix-based structure is regular enough to allow for efficient parallel implementation on the GPUs. By applying a divide-and-conquer approach, we are able to make use of multi-node GPU clusters, resulting in more than an order of magnitude speedup over fastest known Dijkstra-based GPU implementation and a two-fold speedup over a parallel Dijkstra-based CPU implementation.

7.4.5. Analytical models and Optimization for GPUs

Participants: R. Andonov

In [28] we develop a methodology for modeling the energy efficiency of tiled nested-loop codes running on a graphics processing unit (GPU) and use it for energy efficiency optimization. We use the polyhedral model, and we assume that a highly optimized and parametrized version of a tiled nested – loop code, either written by an expert programmer or automatically produced by a polyhedral compilation tool – is given to us as an input. We then model the energy consumption as an analytical function of a set of parameters characterizing the software and the GPU hardware. Our approach develops analytical models based on (i) machine and architecture parameters, (ii) program size parameters as found in the polyhedral model and (iii) tiling parameters, such as those that are chosen by auto-or manual tuners. Our model therefore allows efficient optimization of the energy efficiency with respect to a set of parameters of interest.

7.5. Applications

7.5.1. CAMI: Critical Assessment of Metagenomic Interpretation

Participants: C. Deltel, D. Lavenier, C. Lemaitre, P. Peterlongo, G. Rizk

The interpretation of metagenomes relies on sophisticated computational approaches such as short read assembly, binning and taxonomic classification. All subsequent analyses can only be as meaningful as the outcome of these initial data processing methods [0]. The CAMI initiative aims to evaluate these methods independently, comprehensively and without bias. The goal is to supply users with exhaustive quantitative data about the performance of methods in many relevant scenarios. In 2015, we participate to CAMI within the "assembly" category using the Minia assembly pipeline. Results are provided here: https://data.cami-challenge.org/. For the medium challenge datasets, our assemblies are refered under the identifiers *goofy-wilson* and *fervent-blackwell*.

7.5.2. Assembly and Annotation of Arthropods Genomes

Participants: A. Gouin, F. Legeai, C. Lemaitre

Within a large international network of biologists, GenScale has contributed to various projects for identifying important components such as protein coding or non coding genes involved in the adaptation of major agricultural pests to their environment. We provided the assembly and the annotation of 4 new aphids, 3 parasitic wasps, and improved the assembly of 2 variants of fall army worm by removing unwanted sequences due to heterozygosity [41], [42]. Following specific agreement or policy, these new genomes and annotations are available for a restricted consortium or a large community through the BioInformatics platform for Agro-ecosystems Arthropods (http://bipaa.genouest.org/is). These results, and further analyses led to a better understanding of the biology, evolution and life history traits of *Spodoptera frugiperda* [19], the identiifcation and characterization of new genome of pea aphid symbionts [22] and the identification of differentially expressed genes in the sensory system of Sesamia nonagrioides [21].

7.5.3. Study of the rapeseed genome structure

Participants: D. Lavenier, C. Lemaitre, S. Letort, P. Peterlongo

In collaboration with IGEPP (Institut de Génétique, Environnement et Protection des Plantes), INRA, and through two national projects, PIA Rapsodyn and France-Génomique Polysuccess, we are involved in the genome analysis of several rapeseed varieties. The Rapsodyn project has the ambition to insure long-term competitiveness of the rapeseed production through improvement of the oil yield and reduction of nitrogen inputs during the crop cycle. Rapeseed varieties must thus be selected from genotypes that favor low nitrogen input. DiscoSNP++ is here used to locate new variants among the large panel of rapeseed varieties which have been sequenced during the project. The PolySuccess project aims to answer the following question: how a polyploid, such as the oilseed rape plant, becomes a new species? Oilseed rape (Brassica napus) beeing a natural hybrid between B.rapa and B.oleracea, different genomes of these three species have been sequenced to study their structures. The Minia assembly pipeline provides a fast way to generate contigs that are used for studying gene specificities.

[0] http://www.cami-challenge.org/

SAGE Project-Team

7. New Results

7.1. Numerical algorithms

7.1.1. Introduction to computational linear algebra

Participant: Jocelyne Erhel.

Publications: [23]

Abstract: The book "Introduction to Computational Linear Algebra" presents classroom-tested material on computational linear algebra and its application to numerical solutions of partial and ordinary differential equations. The book is designed for senior undergraduate students in mathematics and engineering as well as first-year graduate students in engineering and computational science.

The text first introduces BLAS operations of types 1, 2, and 3 adapted to a scientific computer environment, specifically MATLAB. It next covers the basic mathematical tools needed in numerical linear algebra and discusses classical material on Gauss decompositions as well as LU and Cholesky's factorizations of matrices. The text then shows how to solve linear least squares problems, provides a detailed numerical treatment of the algebraic eigenvalue problem, and discusses (indirect) iterative methods to solve a system of linear equations. The final chapter illustrates how to solve discretized sparse systems of linear equations. Each chapter ends with exercises and computer projects.

7.1.2. Hybrid algebraic sparse linear solvers

Participants: Jocelyne Erhel, David Imberti.

Grants and projects: EXA2CT 9.2.1 , EoCoE 9.2.2 , C2S@EXA 9.1.2

Publications: in preparation.

Abstract: Sparse linear systems arise in computational science and engineering. The goal is to reduce the memory requirements and the computational cost, by means of high performance computing algorithms. Krylov methods combined with Domain Decomposition are very efficient for both fast convergence and fast computations.

7.1.3. Hastings-Metropolis Algorithm on Markov Chains for Small-Probability Estimation

Participant: Lionel Lenôtre.

Grants: H2MNO4 9.1.1

Publications: [12]

Abstract: Shielding studies in neutron transport, with Monte Carlo codes, yield challenging problems of small-probability estimation. The particularity of these studies is that the small probability to estimate is formulated in terms of the distribution of a Markov chain, instead of that of a random vector in more classical cases. Thus, it is not straightforward to adapt classical statistical methods, for estimating small probabilities involving random vectors, to these neutron-transport problems. A recent interacting-particle method for small-probability estimation, relying on the Hastings-Metropolis algorithm, is presented. It is shown how to adapt the Hastings-Metropolis algorithm when dealing with Markov chains. A convergence result is also shown. Then, the practical implementation of the resulting method for small-probability estimation is treated in details, for a Monte Carlo shielding study. Finally, it is shown, for this study, that the proposed interacting-particle method considerably outperforms a simple Monte Carlo method, when the probability to estimate is small.

7.1.4. A Strategy for the Parallel Implementations of Stochastic Lagrangian Methods

Participant: Lionel Lenôtre.

Grants and projects: H2MNO4 9.1.1

Software: PALMTREE 6.5

Publications: [32]

Abstract: We present some investigations on the parallelization of a stochastic Lagrangian simulation. For the self sufficiency of this work, we start by recalling the stochastic methods used to solve Parabolic Partial Differential Equations with a few physical remarks. Then, we exhibit different object-oriented ideas for such methods. In order to clearly illustrate these ideas, we give an overview of the library PALMTREE that we developed. After these considerations, we discuss the importance of the management of random numbers and argue for the choice of a particular strategy. To support our point, we show some numerical experiments of this approach, and display a speedup curve of PALMTREE. Then, we discuss the problem in managing the parallelization scheme. Finally, we analyze the parallelization of hybrid simulation for a system of Partial Differential Equations. We use some works done in hydrogeology to demonstrate the power of such a concept to avoid numerical diffusion in the solution of Fokker-Planck Equations and investigate the problem of parallelizing scheme under the constraint entailed by domain decomposition. We conclude with a presentation of the latest design that was created for PALMTREE and give a sketch of the possible work to get a powerful parallelized scheme.

7.1.5. About a generation of a log-normal correlated field

Participants: Jocelyne Erhel, Géraldine Pichot.

Grants: HYDRINV 9.3.3 , H2MN04 9.1.1

Software: GENFIELD 6.1

Publications: [18].

Abstract: Uncertainty quantification often requires the generation of large realizations of stationary Gaussian random field over a regular grid.

We compare the classical methods used to simulate the field defined by its covariance function, namely the Discrete Spectral method, the Circulant Embedding approach, and the Discrete Karhunen-Loève approximation. We design and implement a parallel algorithm related to the Discrete Spectral method.

7.2. Numerical models and simulations applied to heat transfer

7.2.1. Small scale modeling of porous media

Participants: Édouard Canot, Salwa Mansour.

Grants: ECOS Sud Chili (ARPHYMAT project) 9.3.2

Software: GLiMuH 6.2

Publications: [13]

Conferences: [20]

Abstract: This study is devoted to the heat transfer between two spherical grains separated by a small gap; dry air is located around the grains and a liquid water meniscus is supposed to be present between them. This problem can be seen as a micro-scale cell of an assembly of solid grains, for which we are looking for the effective thermal conductivity. For a fixed contact angle and according to the volume of the liquid meniscus, two different shapes are possible for the meniscus, giving a “contacting” state (when the liquid makes a true bridge between the two spheres) and a “non-contacting” one (when the liquid is split in two different drops, separated by a thin air layer); the transition between these two states occurs at different times when increasing or decreasing the liquid volume, thus leading to a hysteresis behavior when computing the thermal flux across the domain. We consider also another process where humidity varies, for example during an evaporation or condensation process; in this situation, the shape of the menisci changes a lot, because some liquid bridges may break, and this can strongly affect the effective thermal conductivity. Then, the reorganization of the liquid menisci is predicted, especially their surface area variation; it is an important parameter for a global model of the evaporation phenomenon in wet porous media.

7.2.2. Inverse problem for determining the thermo-physical properties of a porous media

Participants: Édouard Canot, Salwa Mansour.

Grants: HYDRINV 9.3.3

Software: TPIP (6.7)

Publications: [15], [27]

Conferences: [22]

Abstract: This study concerns the inverse problem which consists of the estimation of thermophysical properties of the soil knowing the temperature at few selected points of the domain. In order to solve this inverse problem, we used the least square criterion where we try to minimize the error function between real measures and simulated ones. The coupled system composed of the energy equation together with the three sensitivity boundary initial problems resulting from differentiating the basic energy equation with respect to the soil properties must be solved. To overcome the stiffness of our problem (due to the use of Apparent Heat Capacity method), the high nonlinearity of the coupled system and the problem of large residuals we used the Damped Gauss Newton and Levenberg-Marquardt methods. To take into account uncertainties of the position of the sensors, some constraints have been added to the least square problem. Results are good when the number of sensors is sufficiently large.

7.2.3. Evaporation/Condensation in a wet granular medium: the EWGM model

Participants: Édouard Canot, Salwa Mansour.

Grants: ECOS Sud Chili (ARPHYMAT project) 9.3.2

Software: HeMaTis (6.4)

Publications: [26], [25]

Abstract: The physical model of the HeMaTiS code (6.4) has been completed by a new variant dedicated to the unsaturated case. The pendular regime concerns the special case where a very few quantity of liquid water is contained in a granular medium. The new model involves seven variables and can be considered as a two-phase two-component one; it contains both air and water, this latter component being liquid or gas. Generally, the diffusive transport of humidity in soils is extremely slow, we numerically show that humidity is convected quickly when the medium is subjected to a strong temperature gradient. The key feature of the thermal process is the simultaneous evaporation and condensation of water near a discontinuity of the liquid layout.

7.3. Models and simulations for skew diffusion

7.3.1. Simulating Diffusion Processes in Discontinuous Media: Benchmark Tests

Participant: Géraldine Pichot.

Grants: H2MN04 9.1.1

Software: SBM 6.6

Publications: submitted.

Abstract: We present several benchmark tests for Monte Carlo methods for simulating diffusion in one-dimensional discontinuous media, such as the ones arising the geophysics and many other domains. These benchmarks tests are developed according to their physical, statistical, analytic and numerical relevance. We then perform a systematic study on four numerical methods.

7.3.2. One-dimensional skew diffusions: explicit expressions of densities and resolvent kernel

Participants: Lionel Lenôtre, Géraldine Pichot.

Grants: H2MN04 9.1.1

Publications: [31]

Abstract: The study of skew diffusion is of primary concern for their implication in the modeling and simulation of diffusion phenomenons in media with interfaces. First, we provide results on one-dimensional processes with discontinuous coefficients and their connections with the Feller theory of generators as well as the one of stochastic differential equations involving local time. Second, in view of developing new simulation techniques, we give a method to compute the density and the resolvent kernel of skew diffusions. Explicit closed-form are given for some particular cases.

7.3.3. Algorithms for the simulation of Feller processes

Participant: Lionel Lenôtre.

Grants and projects: H2MNO4 9.1.1 .

Publications: [34].

Abstract: Two new numerical schemes are created for Skew Diffusions processes. Both algorithms rely on a more generic numerical scheme that can be used for any kind of Feller processes. The proof of convergence for this generic numerical scheme is performed.

7.3.4. Theoretical results on multidimensional Skew Diffusions

Participant: Lionel Lenôtre.

Grants and projects: H2MNO4 9.1.1 .

Publications: [33].

Abstract: Some significant results on the distribution of the marginal processes of multidimensional Skew Diffusions are found together with new formula. In addition, totally analytical proofs of some results and algorithms given by A. Lejay are given.

7.4. Models and simulations for flow and transport in porous fractured media

7.4.1. An adaptive sparse grid method for elliptic PDEs with stochastic coefficients

Participant: Jocelyne Erhel.

Grants and projects: HYDRINV 9.3.3 , H2MN04 9.1.1

Publications: [14].

Abstract: The stochastic collocation method based on the anisotropic sparse grid has become a significant tool to solve partial differential equations with stochastic inputs. The aim is to seek a vector of weights and a convenient level of interpolation for the method. The classical approach uses an a posteriori approach on the solution, which causes an additional prohibitive cost.

In this work, we discuss an adaptive approach of this method to calculate the statistics of the solution. It is based on an adaptive approximation of the *inverse* diffusion parameter. We construct an efficient error indicator which is an upper bound of the error on the solution. In the case of unbounded variables, we use an appropriate error estimation to compute suitable weights for the method. Numerical examples are presented to confirm the efficiency of the approach, and to show that the cost is considerably reduced without loss of accuracy.

7.4.2. A global reactive transport model applied to the MoMaS benchmark

Participant: Jocelyne Erhel.

Grants and projects: H2MN04 9.1.1

Software: GRT3D 6.3

Publications: [19].

Abstract: Reactive transport models are very useful for groundwater studies such as water quality, safety analysis of waste disposal, remediation, and so on. The MoMaS group defined a benchmark with several test cases. We present results obtained with a global method and show through these results the efficiency of our numerical model.

7.4.3. About some numerical models for geochemistry

Participant: Jocelyne Erhel.

Grants and projects: H2MN04 9.1.1

Publications: [16], [17].

Abstract: Reactive transport models are very useful to study the fate of contaminants in grounwater. These models couple transport equations with geochemistry equations. In this talk, we focus on precipitation and dissolution chemical reactions, because they induce numerical difficulties.

We consider a set of solute species and minerals, with precipitation occuring when a saturation threshold is reached. A challenge is to detect which minerals are dissolved and which minerals are precipitated. This depends on the total quantities of chemical species. We propose an analytical approach to build a phase diagram, which provides the interfaces between the different possible cases. We illustrate our method with three examples arising from brine media and acid mine drainage.

7.4.4. Power-averaging method to characterize and upscale permeability in DFNs

Participants: Jean-Raynald de Dreuzy, Géraldine Pichot.

Publications: [21].

Abstract: In a lot of geological environments, permeability is dominated by the existence of fractures and by their degree of interconnections. Flow properties depend mainly on the statistical properties of the fracture population (length, apertures, orientation), on the network topology, as well as on some detailed properties within fracture planes. Based on an extensive analysis of 2D and 3D DFNs as well as on reference connectivity structures, we investigate the relation between the local fracture structures and the effective permeability. Defined as the relative weight between the two extreme harmonic and arithmetic means, the power-law averaging exponent gives a compact way to compare fracture network hydraulics. It may further lead to some comprehensive upscaling rules.

SERPICO Project-Team

7. New Results

7.1. Statistical aggregation methods for image denoising and estimation

Participants: Charles Kervrann, Frédéric Lavancier.

We have already proposed a general statistical aggregation method which combines image patches denoised with several commonly-used algorithms [10]. We showed that weakly denoised versions of the input image obtained with standard methods, can serve to compute an efficient patch-based aggregated estimator. In our approach, the Stein's Unbiased Risk Estimator (SURE) is used to evaluate each denoised candidate image patch and to compute the exponential weighted aggregation (EWA) estimator. This year, we adapted this framework (PEWA) to denoise images corrupted by mixed Gaussian-Poisson in 2D fluorescence image sequences.

In this range of work, we have also introduced in [24] a general method to combine estimators in order to produce a better estimate. From a theoretical point of view, we proved that this method is optimal in some sense. It is illustrated on standard statistical problems in parametric and semi-parametric models where the averaging estimator outperforms the initial estimators in most cases. As part of an on-going work, we are applying this method to improve patch-based image denoising algorithms.

Reference: [24]

Collaborators: Paul Rochet (Laboratoire de Mathématiques Jean Leray (LMJL), University of Nantes).

7.2. Image deconvolution algorithms for tagged-RNA and gene localization in live yeast

Participant: Charles Kervrann.

In fluorescence microscopy, the image quality is limited by out-of-focus blur and high noise. Traditionally, image deconvolution is needed to estimate a good quality version of the observed image. The result of deconvolution depends heavily on the choice of the regularization term. The regularization functional should be designed to remove noise while retaining the image structure. In this study, we investigated non quadratic regularization terms to preserve fine details of underlying structures and we studied appropriate optimization algorithms. The deconvolution method has been especially dedicated for 3D high-precision gene localization in cell nuclei [47]. For illustration, tagged gene (green marker) and tagged nucleoporins/nuclear periphery (red marker) are shown in Fig. 3 . A noisy and blurred image can affect the nuclear membrane estimation and gene detection and, consequently, the computed related distances.

Collaborators: Giovanni Petrazzuoli (Inserm U944, CNRS UMR 9212, Hôpital Saint-Louis, Paris), Catherine Dargemont (Inserm U944, CNRS UMR 9212, Hôpital Saint-Louis, Paris), Jean Salamero (UMR 144 CNRS-Institut Curie, PICT-IBiSA).

7.3. Estimation of the reference point giving the most uniform angular distribution

Participants: Thierry Pécot, Patrick Bouthemy, Charles Kervrann.

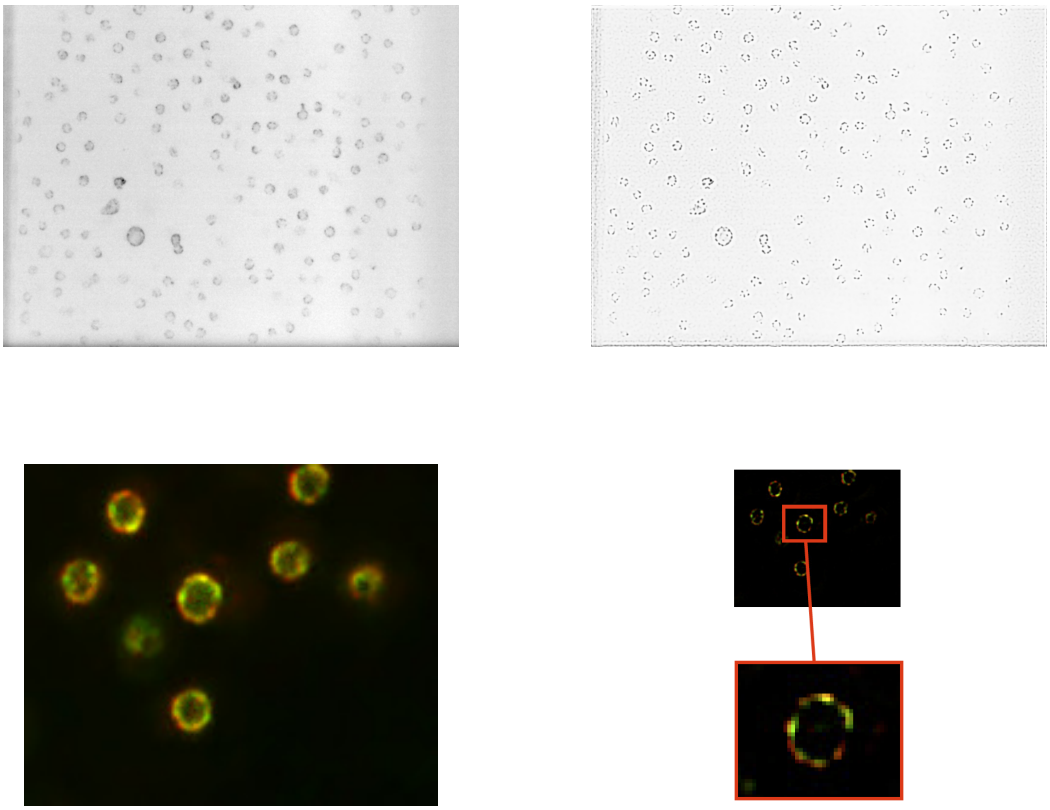


Figure 3. Deconvolution of 3D image depicting tagged gene and tagged nucleoporins/nuclear periphery. First row: deconvolution (right) of a tagged nucleoporin image (left). Second row: blurred image of tagged gene and nucleoporins (left) and zoom-in view of the deblurred image.

Rab11 proteins are trafficking from the Endosomal Recycling Compartment (ERC) to locations in the cell membrane where they eventually fuse. In this study, we assume that the Rab11 positive membranes are uniformly distributed around the ERC at the cell membrane. To test this hypothesis, we estimate the angular distribution of Rab11 positive membranes from several image sequences acquired with a TIRF microscope at the cell membrane level by considering all the points located in the cell as a reference point. We then compute the entropy of angular distribution for each point and estimate the ERC location as the reference point that gives the maximum entropy for the angular distribution (see Fig. 4). These results are very close to the ERC locations manually annotated by experts.

Collaborators: Jean Salamero (UMR 144 CNRS-Institut Curie, PICT-IBiSA),
Jérôme Boulanger (UMR 144 CNRS-Institut Curie),
Liu Zengzhen (UMR 144 CNRS-Institut Curie).

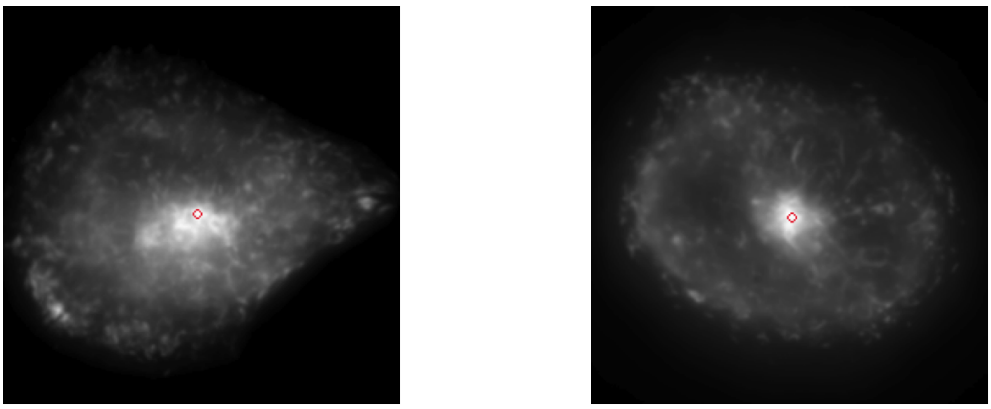


Figure 4. Estimation of reference points (red circles) for Rab11 traffic comparison, registration and quantification.

7.4. Modeling and estimation of protein release and diffusion in TIRFM

Participants: Antoine Basset, Charles Kervrann, Patrick Bouthemy.

We have pursued our work on membrane dynamics, still following a local approach in space and time. We have proposed a new model to account for the full behavior of cargo transmembrane proteins during the vesicle fusion to the plasma membrane at the end of the exocytosis process (see Fig. 5). It combines release and diffusion steps. The former is represented by an exponential decay to account for a continuous release of the proteins from the vesicle to the plasma membrane. We can relax the usual point source assumption, and we name our model the "Small-extent Source with Exponential Decay release" (SSED). An iterative minimization method is used to estimate simultaneously both biophysical parameters, i.e., the release rate and the diffusion coefficient, for every active vesicle detected in the total internal reference fluorescence microscopy (TIRFM) image sequence. Quantitative evaluation has demonstrated the efficiency of the method, which has also allowed us to exhibit differences in the behaviors of Transferrin receptor (TfR) and Langerin proteins.

Collaborators: Jean Salamero (UMR 144 CNRS-Institut Curie, PICT-IBiSA),
Jérôme Boulanger (UMR 144 CNRS-Institut Curie).

7.5. Counting-based particle flux estimation for traffic analysis in live cell imaging

Participants: Thierry Pécot, Charles Kervrann.

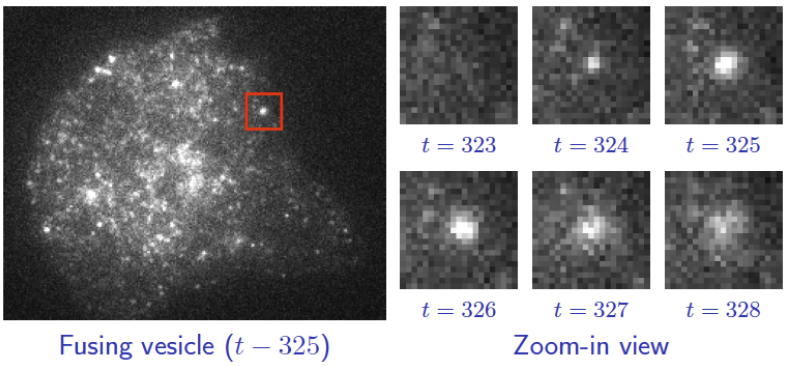


Figure 5. Left: Fusing vesicle (frame in red) in a TIRFM (UMR 144 CNRS-Institut Curie, PICT-IBiSA) sequence (frame 325, 50ms/frame). Right: Zoom-in view of the temporal evolution of the fusing vesicle.

In this study, we have proposed an original traffic analysis approach based on the counting of particles from frame to frame. Object tracking methods or optical flow methods are generally considered to analyze the dynamic contents of intracellular video-microscopy. The suggested method lies between these two well-known approaches. Instead of tracking each moving particle, we estimate fluxes of particles between predefined and adjacent regions. Our three-step counting-based approach is as follows:

- The cell is uniformly partitioned into fixed-size and fixed-shape regions.
- The moving particles are automatically detected using an appropriate algorithm.
- The fluxes are estimated with sparse constraints from an image pair at each time step from temporal variations of the number of particles in each region of the uniform tessellation. Except for some trivial cases, the flux estimation is actually an ill-posed problem and additional constraints are necessary to find the optimal solution.

The problem is formulated as the minimization of a global cost function and the approach allows us to process image sequences with a high number of particles and a high rate of particle appearances and disappearances. We studied the influence of object density, image partition scale, motion amplitude and particle appearances/disappearances in a large variety of simulations. The potential of the method has been demonstrated on real image sequences showing GFP-tagged Rab6 trafficking in confocal microscopy.

Reference: [26]

Collaborators: Jean Salamero (UMR 144 CNRS-Institut Curie, PICT-IBiSA),
Jérôme Boulanger (UMR 144 CNRS-Institut Curie).

7.6. Tracking of astral microtubules at the cell cortex

Participants: Frédéric Logé-Munerel, Thierry Pécot, Antoine Basset, Charles Kervrann.

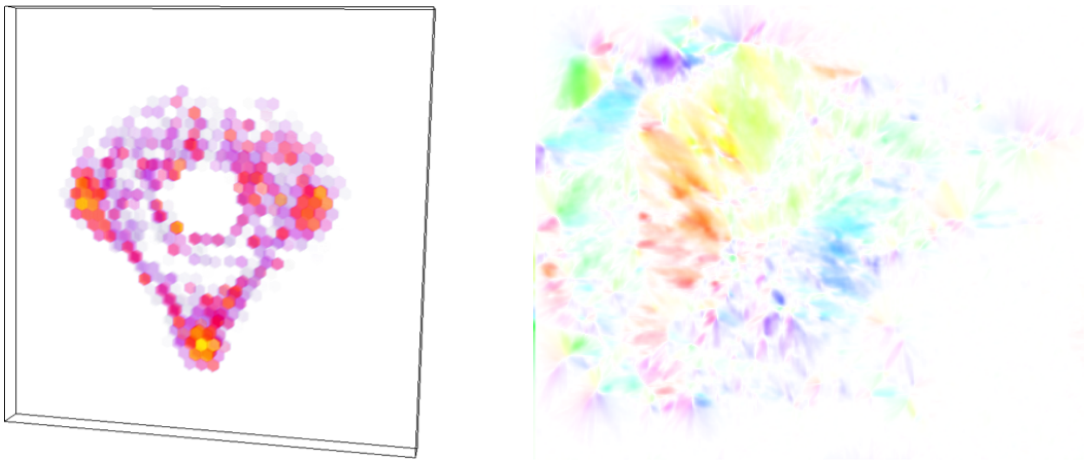


Figure 6. Left: Spatial distribution of GFP-tagged Rab6 vesicle numbers estimated when considering a regular 3D tesselation. Right: Estimation of directional/intensity flows of actin filaments for an image sequence acquired in TIRF microscopy (UMR 144 CNRS-Institut Curie, PICT-IBiSA).

In this study, we are currently interested in the influence of the mechanical properties of astral microtubules in the centering mechanisms of the mitotic spindle, giving it a robust positioning. In their previous studies, the CeDRE group (IGDR Rennes) identified two subpopulations of astral microtubules that either push or pull the cell cortex. To better understand these mechanisms, they acquired image sequences at the cortex level where astral microtubules extremities come to exert forces. In order to characterize the two subpopulations of astral microtubules during the mitosis in the unicellular embryos of C. Elegans life span, that is the period during which the microtubule is touching the cell cortex, has to be measured for every single microtubule. A short life span corresponds to a pulling force and a long life span corresponds to a pushing force. Detecting and tracking microtubules at the cell cortex has to be done to collect these measures. This year, F. Logé-Munerel (internship Master 1, supervisors: T. Pécot and C. Kervrann) improved the analysis workflow and calibrated the parameters of the algorithms to successfully track the microtubules. This workflow is composed of the ND-SAFIR denoising algorithm [4], the ATLAS detection algorithm [12] and the ASTRE tracking algorithm [56]. The experimental results are currently compared with results obtained by the CeDRE group using the U-track platform [50] (see Fig. 7).

Collaborators: Jacques Pécréaux and Hélène Bouvrais (CeDRE group, IGDR Rennes, CNRS UMR 6290).

7.7. Correlation-based method for membrane diffusion estimation during exocytosis in TIRFM

Participants: Ancageorgiana Caranfil, Antoine Basset, Charles Kervrann.

The dynamics of the plasma membrane of the cell is not fully understood yet; one of the crucial aspects to clarify is the diffusion process during exocytosis. Several observation methods exist, including TIRFM (Total Internal Reflection Fluorescence Microscopy), that has successfully been used to determine the successive steps of exocytosis. However, computing characteristic values for plasma membrane dynamics is problematic, as the experimental conditions have a strong influence on the obtained data and a global model cannot be determined. The goal of this study was to build a correlation-like method to estimate local diffusion parameters in TIRFM images. Using a correlation approach similar to TICS (Temporal Image Correlation Spectroscopy) with an adapted local model, we have developed a novel correlation-based method to estimate the diffusion

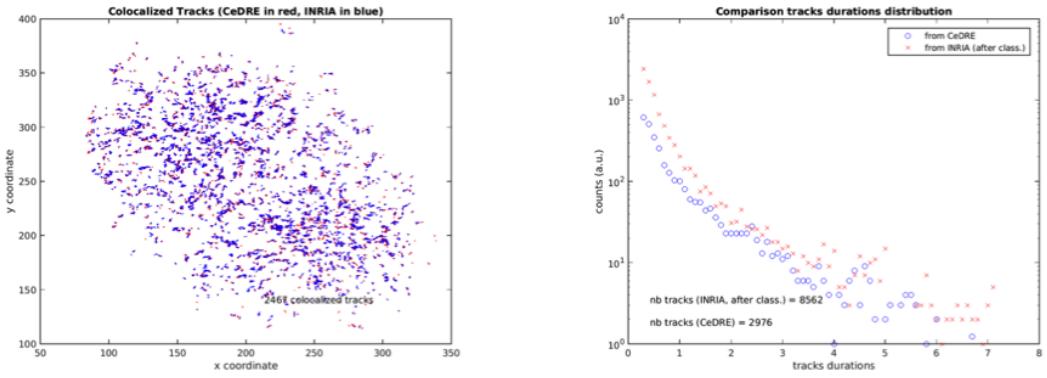


Figure 7.

Microtubule extremities detection and tracking in fluorescence microscopy (embryo of C. Elegans, IGDR - Institute of Genetics and Developmental biology of Rennes, CNRS UMR 6290).

coefficient for every diffusion event in TIRFM images. We turned the non-linear model of the TICS method into a linear one, and made it rely on less parameters than the other estimation methods. Results are excellent for sequences with a good signal-to-noise ratio (see Fig. 8); however, time and space dependencies are introduced with the presence of moderate-to-strong image noise. Although only synthetic images have been used so far, studies of real-life TIRFM images are forthcoming, along with refinements to make the method robust to noise.

Collaborators: Perrine Paul-Gilloteaux and Francois Waharte (UMR 144 CNRS-Institut Curie, PICT-IBiSA).

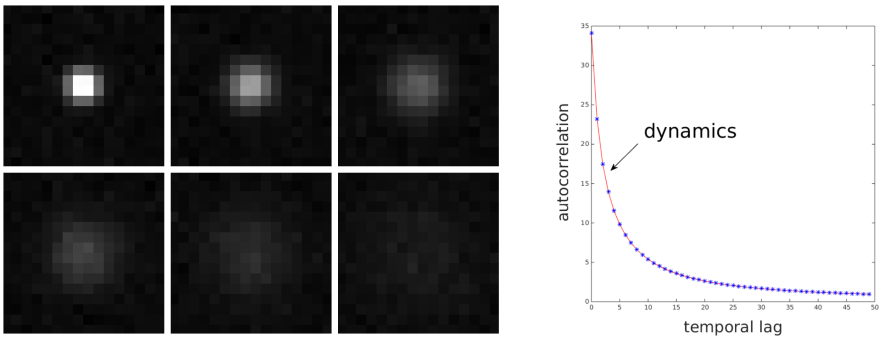


Figure 8. Left: first six instances of a TIRFM image sequence showing a diffusion event. Right: the correlation-based method is applied on the TIRFM sequence; both the computed values of the autocorrelation, for different values of the temporal lag, and the fitting function for these values are represented.

7.8. Co-localization between proteins : testing procedure and generative models

Participants: Frédéric Lavancier, Thierry Pécot, Charles Kervrann.

In the context of bioimaging, co-localization refers to the detection of emissions from two or more fluorescent molecules within the same pixel of the image. This approach enables to quantify the protein-protein interactions inside the cell, just at the resolution limit of the microscope. In statistics, this amounts to characterizing the joint spatial repartition and the spatial overlap between different fluorescent labels. An illustration of the co-localization of green (Langerin protein) and red (Rab11 GTPase protein) fluorescence is shown in Fig. 9 (the images were segmented by applying the ATLAS algorithm [12]). In our framework, the spatial repartition of proteins in the same cell is modeled by a union of random balls, possibly overlapping, and a Gibbs interaction is introduced to take into account the possible interaction between the two co-expressed proteins. A simulation algorithm is described and an inference procedure, based on the Takacs-Fiksel method, is proposed to estimate the interaction parameter. This estimation allows us to determine the presence of co-localization and to quantify the degree of interactions. On the other hand, this model can be used as a generator for synthesized images of co-localized proteins, in a view to assess testing procedures as the one explained below.

In an on-going project, we are developing a non-parametric testing procedure for co-localization. It is mainly based on the overlap area, corresponding to yellow spots as displayed in the right-hand side image of Fig. 9 . Our first experiments on synthesized images showed that our procedure is more powerful than all existing methods to detect co-localization. Moreover this testing procedure turns out to be robust to different shapes and sizes of objects segmented by any competitive algorithm.

Reference: [36]

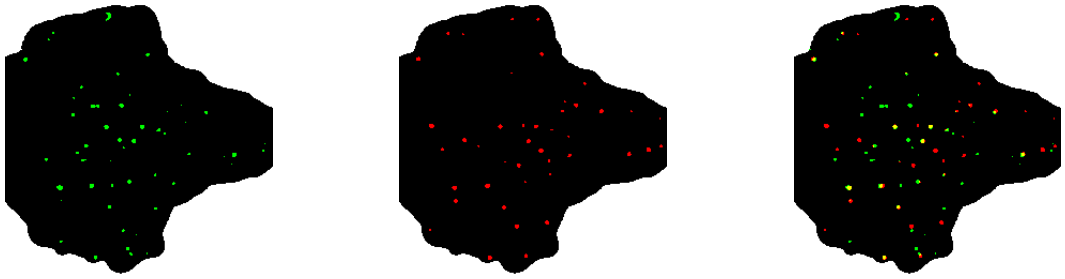


Figure 9. M10 cell showing Langerin proteins (left, in green) and Rab11 GTPase proteins (middle, in red). Right: superposition of the two previous images resulting in some possible yellow spots (co-expression of proteins within the same pixel).

7.9. Classification of diffusion dynamics from particle trajectories

Participants: Vincent Briane, Charles Kervrann.

In this study, we are currently interested in describing the dynamics of particles inside live cell. We assume that the motions of particles follow a certain class of random process: the diffusion processes. We have proposed a statistical method able to classify the motion of the observed trajectories into three groups: “confined”, “directed” and “free diffusion” (namely Brownian motion). This method is an alternative to the commonly used Mean Square Displacement (MSD) analysis. We assessed our procedure on both simulations and real cases; an example of confined diffusion is the Ornstein-Uhlenbeck process while an example of directed diffusion is the Brownian motion with constant drift. The method is currently applied to investigate membrane trafficking (Rab11/Langerin (see Fig. 10) and Rab11/TfR protein sequences) using the following procedure:

1. Tracking of particles with any competitive algorithm.
2. Statistical test /classification applied on tracks longer than ten time points.
3. Estimation of diffusion parameters (e.g. drift, diffusion, ...).

Each trajectory is labelled with the most likely process and the parameters of the underlying process are estimated. Future work will concern the detection of change of motion dynamic over time. Some results of our test on the Langerin protein sequence are shown in Fig. 10 .

Collaborator: Myriam Vimond (ENSAI Rennes).

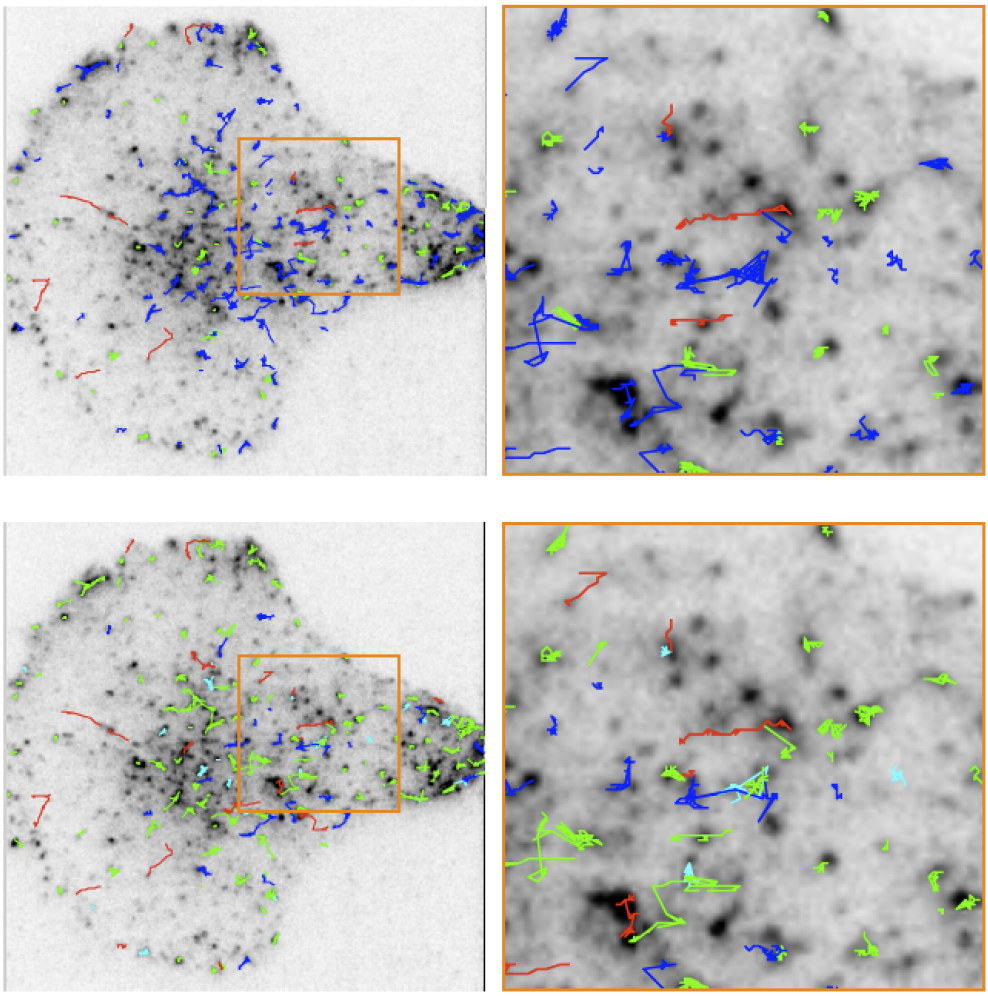


Figure 10. Labelling of the dynamics of trajectories on the Langerin protein sequence (Courtesy of UMR 144 CNRS-Institut Curie and PICT IBiSA). We display only the trajectories appearing on the first 100 frames. The color code is red for directed Brownian, green for Ornstein-Uhlenbeck, blue for Brownian, cyan for motionless. Top panel is labelled with our test, bottom panel with the MSD method.

7.10. Inference for spatial Gibbs point processes

Participant: Frédéric Lavancier.

Gibbs point processes are popular and widely used models in spatial statistics to describe the repartition of points or geometrical structures in space. They initially arose from statistical physics where they are models for interacting particles. They are now used in as different domains as astronomy, biology, computer science, ecology, forestry, image analysis and materials science. Assuming a parametric form of the Gibbs interaction, the natural method to estimate the parameters is likelihood inference. Since its first use in the 80's, this method is conjectured to be consistent and efficient. However the theoretical properties of maximum likelihood for Gibbs point processes remain largely unknown. In [39], we have partly solved this 30 years old conjecture by proving the consistency of the likelihood procedure for a large class of Gibbs models. As important examples, we deduced the consistency of the maximum likelihood estimator for all parameters of the Strauss model, the hardcore Strauss model, the Lennard-Jones model and the area-interaction model, which are commonly used models in practice.

A practical issue of likelihood estimation yet is that this method depends on an intractable normalizing constant that has to be approximated by simulation. To avoid this problem, other methods of estimation have been introduced, including pseudo-likelihood estimation. The theoretical properties of the pseudo-likelihood method are fairly well known in the case of finite-range Gibbs interactions. However, this setting rules out some major Gibbs models as the Lennard-Jones model. In [15], we have extended the pseudo-likelihood procedure to infinite range Gibbs interactions and proved its consistency and its asymptotic normality.

References: [15], [39]

Collaborators: David Dereudre (Laboratoire Paul Painlevé (UMR 8524), University of Lille 1),
Jean-François Coeurjolly (Laboratoire Jean Kutzmann, University of Grenoble).

7.11. Statistical aspects of Determinantal Point Processes

Participant: Frédéric Lavancier.

Determinantal point processes (DPPs) have been introduced in their general form by Macchi (1975) and have been extensively studied from a probabilistic point of view in the 2000's (one of the main reason being their central role in random matrix theory). In [23], we have demonstrated that DPPs provide useful models for the description of spatial point pattern datasets where nearby points repel each other. We have exploited the appealing probabilistic properties of DPPs to develop parametric models, where the likelihood and moment expressions can be easily evaluated and realizations can be quickly simulated. We have discussed how statistical inference is conducted using the likelihood or moment properties of DPP models, and we provided freely available software for simulation and statistical inference.

In [13], we have addressed the question of how repulsive a stationary DPP can be, in order to assess the range of practical situations this promising class of models may model. We determine the most repulsive DPP (in some sense) and we introduce new parametric families of stationary DPPs that can cover a large range of DPPs, from the stationary Poisson process (the case of no interaction) to the most repulsive DPP. Some theoretical aspects of inference for stationary DPPs are tackled in [37] and [38]. In the former study we have established the Brillinger mixing property of stationary DPPs, a first important step toward asymptotic inference. In the latter contribution, we have exploited this result to deduce the consistency and asymptotic properties of contrast estimators for stationary DPPs.

References: [23], [13], [37], [38]

Collaborators: Christophe Ange Napoléon Biscio (LMJL, University of Nantes),
Jesper Møller (Department of Mathematical Sciences, Aalborg University, Denmark),
Ege Rubak (Department of Mathematical Sciences, Aalborg University, Denmark).

7.12. Modelling aggregation and regularity in spatial point pattern datasets

Participant: Frédéric Lavancier.

In the spatial point process literature, analysis of spatial point pattern datasets are often classified into three main cases: i/ regularity (or inhibition or repulsiveness), modelled by Gibbs point processes, hard core processes like Matern hard core models, and determinantal point processes; ii/ complete spatial randomness, modelled by Poisson point processes; iii/ aggregation (or clustering), modelled by Poisson cluster processes and Cox processes. For applications the classification i/-iii/ can be too simplistic, and there is a lack of useful spatial point process models with, loosely speaking, aggregation on the large scale and regularity on the small scale. For instance, we may be interested in such a model for the repartition of the centres of vesicles in a cell, that exhibit some spatial clustering at large scales while having a minimal distance between them.

In [22], we have considered a dependent thinning of a regular point process with the aim of obtaining aggregation on the large scale and regularity on the small scale in the resulting target point process of retained points. Various parametric models for the underlying processes are suggested and the properties of the target point process are studied. Simulation and inference procedures have been discussed when a realization of the target point process is observed, depending on whether the thinned points are also observed or not. Some typical simulations of the target processes are shown in Fig. 11 .

Reference: [22]

Collaborator: Jesper Møller (Department of Mathematical Sciences, Aalborg University, Denmark).

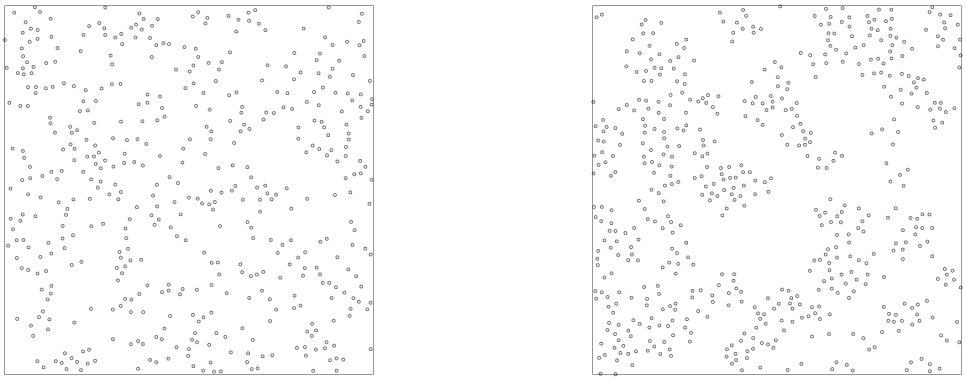


Figure 11. Examples of simulations with aggregation on the large scale and regularity on the small scale.

7.13. Retracing and registration for Correlative light-electron microscopy (CLEM)

Participants: Bertha Mayela Toledo Acosta, Patrick Bouthemy, Charles Kervrann.

Correlative light-electron microscopy (CLEM) enables to relate cell dynamics visualized in light microscopy (LM) with cell structure provided by electron microscopy (EM) for a better understanding of cell mechanisms. Registration of LM and EM modalities is then a timely, important but difficult open problem, which still requires some manual assistance. LM and EM images are indeed of very different size, spatial resolution, field of view, and appearance. We have investigated an original automated approach for the retracing-and-registration stage of the overall CLEM workflow (see Fig. 12). Pairing between the LM region of interest (ROI) and the corresponding EM patch relies on a common representation for both images, based on the LoG (Laplacian of Gaussian) transform with an adaptive associated scale (or blurring). We exploit histograms of the LoG values or histograms supplied by the LDP (Local Directional Pattern) texture descriptor, with associated histogram distances, to solve the EM patch search issue. The search step supplies a pre-registration, which is

refined by the estimation of an affine motion model to overlay the EM image onto the LM image around the ROI. Preliminary results on real CLEM images provided by UMR 144 CNRS-Institut Curie demonstrated the interest and efficiency of the proposed method.

Collaborators: Perrine Paul-Gilloteaux and Xavier Heiligenstein (UMR 144 CNRS-Institut Curie).

7.14. Denoising and compensation of the missing wedge in cryo electron tomography

Participants: Emmanuel Moebel, Charles Kervrann.

In this study, we have addressed two important issues in cryo electron tomography (CET) images: the low signal-to-noise ratio and the presence of a missing wedge (MW) of information in the spectral domain. Indeed, according to the Fourier slice theorem, limited angle tomography results into an incomplete sampling of the Fourier domain. Therefore, the Fourier domain is separated into two regions: the known spectrum (KS) and the unknown spectrum, the latter having the shape of a missing wedge (see Fig. 13). The proposed method tackles both issues jointly, by iteratively applying a denoising algorithm in order to fill up the MW, and proceeds as follows:

1. Excitation step: Add noise into the MW.
2. Denoising step: Apply a patch-based denoising algorithm.
3. Repeat steps 1 and 2, by keeping KS constant through the iterations.

The excitation step is used to randomly initialize the coefficients of the MW, whereas the denoising step acts as a spatial regularization. The employed denoising algorithm, which exploits the self-similarity of the image, filters out coefficient values which are dissimilar to KS, thereby keeping similar ones. By iterating these steps, we are able to diffuse the information contained in KS into the MW.

An application example on experimental data can be seen on Fig. 13 , which shows the data in both spectral and spatial domain. The data contains a spherical gold particle, deformed by MW induced artifacts: elongation of the object, side- and ray-artifacts. From the residue image it can be seen that noise and MW artifacts have been reduced, while preserving the details of the image. Experiments are being performed to verify if particle detection and alignment are enhanced by using the method as a pre-processing step.

Collaborators: Damien Larivière (Fondation Fourmentin-Guilbert),
Julio Ortiz (Max-Planck Institute, Martinsried, Germany).

7.15. Algorithms for row registration to improve quality of Tissue MicroArray (TMA) images

Participants: Hoai Nam Nguyen, Charles Kervrann.

Row jittering is a common problem arising in medical imaging devices such as CT (Computer Tomography) and MRI (Magnetic Resonance Imaging) scanners due to errors of synchronization during image acquisition process. On scanners designed and developed by Innopsys, the problem becomes more challenging mainly because the pixel displacement is non constant along each row (Fig. 14) and possibly sub-pixel (i.e. non integer translation). To overcome this drawback, we first proposed a window-based algorithm to approximate the translation at each pixel by selecting the value that best minimizes a matching criteria over a finite set of possible sub-pixel translations. We obtained satisfying results with this method on real data with fast computation time (see Fig. 14). Furthermore, this matching criteria has been considered as a data fidelity term and was combined to a regularization term to promote a smooth solution and correct small artifacts which were not removed with the window-based method. To minimize the energy functional, we have adopted the quadratic relaxation technique and proximal method. This algorithm is slower and is initalized by the window-based algorithm to produce very encouraging results and elimination of all undesirable artifacts (see Fig. 14).

Collaborators: Vincent Paveau and Cyril Cauchois (Innopys).

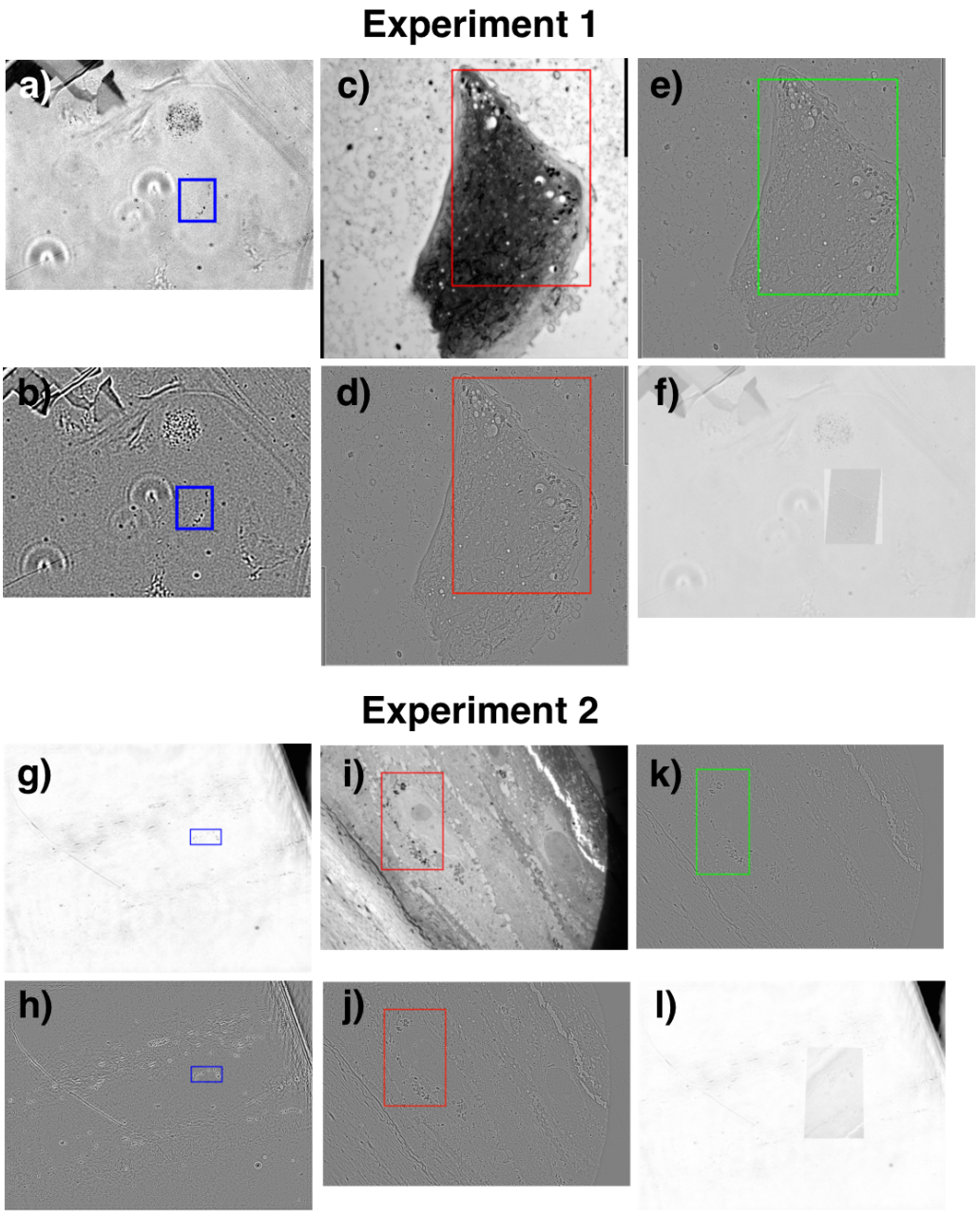


Figure 12. CLEM experiment #1: a) LM image with Region of Interest (ROI) framed in blue; b) same ROI delineated in the LoG-LM image; c) ground-truth location of the corresponding EM patch framed in red; d) the same but in the LoG-EM image; e) selected patch (SP) in the LoG-EM image in green; f) overlay (after registration) of the (decimated) EM image on the LM image around the ROI. CLEM experiment #2: g) LM-ROI in blue; h) LoG-LM-ROI; i) EM-GT in red; j) LoG-EM-GT; k) LoG-EM selected patch; l) Overlay of EM on LM around ROI.

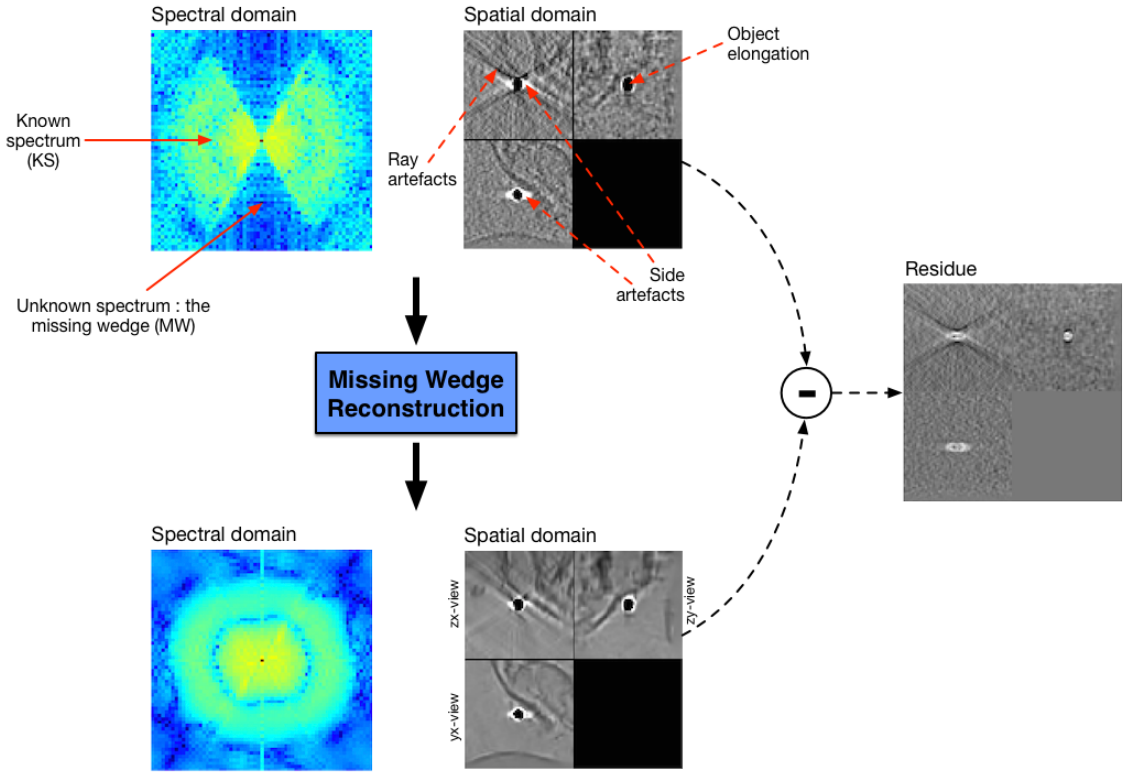


Figure 13. Experimental result of denoising and compensation of the missing wedge in cryo electron tomography.

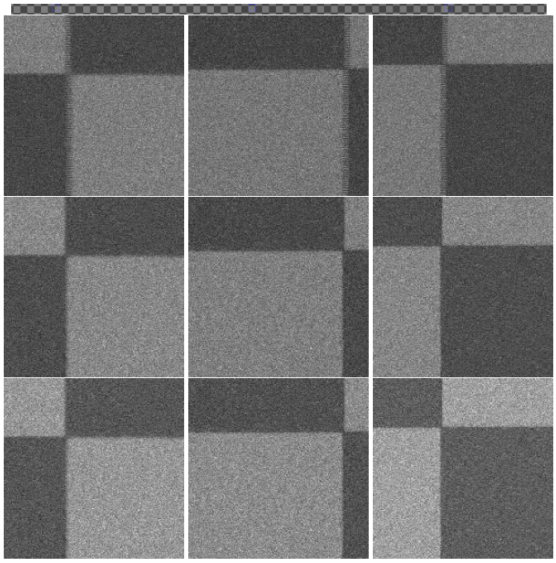


Figure 14. Illustration of the two-row registration algorithms. First row : full width input image (by courtesy of Innopsys). Second row : zooms on input image (blue boxes). Third row : corrected with window-based algorithm. Fourth row : corrected with variational method.

7.16. Robust motion model selection

Participants: Patrick Bouthemy, Bertha Mayela Toledo Acosta.

Parametric motion models are commonly used in image sequence analysis for different tasks. A robust estimation framework is usually required to reliably compute the motion model. However, choosing the most appropriate model in that estimation context is still an open issue. Indeed, penalizing the model complexity while maximizing the size of the inlier set may be contradictory. In this study, we proposed a robust motion model selection method which relies on the Fisher statistic. We also derived an interpretation of it as a robust C_P-Mallows criterion. The resulting criterion is straightforward to compute and explicitly involves the aforementioned trade-off between maximizing the size of the inlier set and minimizing the complexity (i.e., the number of parameters) of the selected motion model. We have conducted a comparative experimental evaluation on synthetic and real image sequences demonstrating that our criterion outperforms the RBIC criterion.

Collaborator: Bernard Delyon (IRMAR Rennes).

7.17. Anomaly detection in crowded scenes

Participants: Juan Manuel Perez Rua, Antoine Basset, Patrick Bouthemy.

We have defined an original motion-based method to detect and localize abnormal events in videos of crowded scenes. The algorithm relies on so-called labeled affine flows, involving both affine motion types and affine velocity vectors, and on view-based crowd motion classes. At every pixel the crowd motion class is inferred from the affine motion model selected among a set of candidate models estimated over a collection of windows. Then, the image is subdivided in blocks where local crowd motion class histograms weighted by the affine motion vector magnitudes are computed. They are block-wise compared to histograms of normal behaviors with a combined distance. More specifically, we introduce the so-called local outlier factor (LOF) to detect anomalous blocks. LOF is a local flexible measure of the relative density of data points in a feature space,

here the space of crowd motion class histograms. By thresholding the LOF value, we can detect an abnormal event in a given block at a given time. Comparative experiments on several real datasets demonstrated that our method is competitive with methods relying on far more elaborated models and exploiting both appearance and motion, while yielding superior performance over motion-based anomaly detection methods.

7.18. Occlusion detection in image sequences

Participants: Juan Manuel Perez Rua, Patrick Bouthemy.

The problem of localizing occlusions between consecutive frames of a video is important but rarely tackled on its own. In most works, it is tightly interleaved with the computation of accurate optical flows, which leads to a delicate chicken-and-egg problem. With this in mind, we proposed a novel approach to occlusion detection where visibility or not of a point in next frame is formulated in terms of visual reconstruction. The key issue is now to determine how well a pixel in the first image can be "reconstructed" from co-located colors in the next image. We first exploited this reasoning at the pixel level with a new detection criterion. Contrary to the ubiquitous displaced-frame-difference, the proposed alternative does not critically depend on a pre-computed, dense displacement field, while being shown to be more effective. We then leveraged this local modeling within an energy-minimization framework that delivers occlusion maps. An easy-to-obtain collection of parametric motion models is exploited within the energy to provide the required level of motion information. Our approach outperforms state-of-the-art detection methods on the challenging MPI Sintel dataset.

Collaborators: Tomas Crivelli and Patrick Pérez (Technicolor).

VISAGES Project-Team

7. New Results

7.1. Image Computing: Detection, Segmentation, Registration and Analysis

7.1.1. Symmetric Block-Matching Registration for the Distortion Correction of Echo-Planar Images

Participants: Renaud Hédouin, Olivier Commowick, Elise Bannier, Christian Barillot.

We introduce a new approach to correct geometric and intensity distortion of Echo Planar Images (EPI) from images acquired with opposite phase encoding directions. A new symmetric block-matching registration algorithm has been developed for this purpose relying on new adapted transformations between blocks and a symmetric optimization scheme to ensure an opposite symmetric transformation. Our results show the ability of our algorithm to robustly recover EPI distortion while obtaining sharper results than the popular TOPUP algorithm [24], [34].

7.1.2. Quantitative analysis of T2/T2 relaxation time alteration*

Participants: Benoit Combès, Anne Kerbrat, Olivier Commowick, Christian Barillot.

T2 and T2* relaxometric data becomes a standard tool for the quantitative assessment of brain tissues and of their changes along time or after the infusion of a contrast agent. Being able to detect significant changes of T2/T2* relaxation time is an important issue. Generally, such a task is performed by comparing the variability level in the regions of interest to the variability in the normal appearance white matter. However, in the case of T2 and T2* relaxometry, this solution is highly problematic. Indeed the level of noise in the normal appearance white matter is significanlty smaller than the level of noise in more intense region (e.g. MS lesions). Our aim is to provide a Bayesian analysis of T2/T2* relaxometry estimation and alteration. More specifically, we build posterior distributions for the relaxation time and the relaxation offset by elucidating the dedicated Jeffreys priors. Then the resulting posterior distributions can be evaluated using a Monte Carlo Markov Chain algorithm. Such an analysis has three main advantages over the classical estimation procedure. First it allows in a simple way to compute many estimators of the posterior including the mode, the mean, the variance and confidence intervals. Then, it allows to include prior information. Finally, becauce one can extract confidence interval from the posterior, testing properly whether the true relaxometry time is included within a certain range of value given a confidence level is simple.

7.1.3. MRI quantitative imaging: Myelin Water Fraction (MWF) quantification in Multiple Sclerosis

Participants: Olivier Commowick, Elise Bannier, Christian Barillot.

Multi-echo T2 relaxometry is potentially a relevant imaging method for MWF quantification in the study of multiple sclerosis (MS). However, to ensure accurate estimation, a large number of echoes are still required that can drive to very long acquisitions. In practice, 32 echo times ranging from 10 ms to 320 ms and an echo spacing (ESP) of 10 ms are used1. Analysis of the decay curve of the consecutive echoes allows the estimation of the T2 spectrum. The proposed approach makes use of recent spatial regularization methods for MWF estimation from clinically compatible acquisitions (typically 11 echoes acquired within 6 minutes). The algorithms were evaluated on both synthetic and clinical data. This work was done during the internship of Lucas Soustelle [32], [29].

7.1.4. Classification of Multiple Sclerosis Lesions using Adaptive Dictionary Learning

Participants: Hrishikesh Deshpande, Pierre Maurel, Christian Barillot.

This work presents a sparse representation and an adaptive dictionary learning based method for automated classification of Multiple Sclerosis (MS) lesions in Magnetic Resonance (MR) images. Manual delineation of MS lesions is a time-consuming task, requiring neuroradiology experts to analyze huge volume of MR data. This, in addition to the high intra- and inter-observer variability necessitates the requirement of automated MS lesion classification methods. Among many image representation models and classification methods that can be used for such purpose, we investigate the use of sparse modeling. In the recent years, sparse representation has evolved as a tool in modeling data using a few basis elements of an over-complete dictionary and has found applications in many image processing tasks including classification. We propose a supervised classification approach by learning dictionaries specific to the lesions and individual healthy brain tissues, which include White Matter (WM), Gray Matter (GM) and Cerebrospinal Fluid (CSF). The size of the dictionaries learned for each class plays a major role in data representation but it is an even more crucial element in the case of competitive classification. Our approach adapts the size of the dictionary for each class, depending on the complexity of the underlying data. The algorithm is validated using 52 multi-sequence MR images acquired from 13 MS patients. The results demonstrate the effectiveness of our approach in MS lesion classification.

This work has been published in the journal of Computerized Medical Imaging and Graphics, Elsevier, 2015 [15]. Part of this work is published as a conference paper in ISBI 2015 [22].

7.1.5. Robust Detection of Multiple Sclerosis Lesions

Participants: Yogesh Karpate, Olivier Commowick, Christian Barillot.

Multiple sclerosis (MS) is a disease with heterogeneous evolution among the patients. Quantitative analysis of longitudinal Magnetic Resonance Images (MRI) provides a spatial analysis of the brain tissues which may lead to the discovery of biomarkers of disease evolution. Better understanding of the disease will lead to a better discovery of pathogenic mechanisms, allowing for patient-adapted therapeutic strategies. To characterize MS lesions, we have proposed two new approaches. The first one consists in a novel paradigm to detect white matter lesions based on a statistical framework [26]. It aims at studying the benefits of using multi-channel MRI to detect statistically significant differences between each individual MS patient and a database of control subjects. This framework consists in two components. First, intensity standardization is conducted to minimize the inter-subject intensity difference arising from variability of the acquisition process and different scanners. The intensity normalization maps parameters obtained using a robust Gaussian Mixture Model (GMM) estimation not affected by the presence of MS lesions. The second part studies the comparison of multi-channel MRI of MS patients with respect to an atlas built from the control subjects, thereby allowing us to look for differences in normal appearing white matter, in and around the lesions of each patient. Experimental results demonstrate that our technique accurately detects significant differences in lesions consequently improving the results of MS lesion detection.

Then we have presented an automatic algorithm for the detection of multiple sclerosis lesions (MSL) from multi-sequence magnetic resonance imaging (MRI) [25]. We built a probabilistic classifier that can recognize MSL as a novel class, trained only on Normal Appearing Brain Tissues (NABT). Patch based intensity information of MRI images is used to train a classifier at the voxel level. The classifier is in turn used to compute a probability characterizing the likelihood of each voxel to be a lesion. This probability is then used to identify a lesion voxel based on simple Otsu thresholding. The proposed framework was evaluated on 16 patients and our analysis reveals that our approach is well suited for MSL detection and outperforms other benchmark approaches.

7.2. Image processing on Diffusion Weighted Magnetic Resonance Imaging

7.2.1. Interpolation and Averaging of Multi-Compartment Model Images

Participants: Renaud Hédouin, Olivier Commowick, Christian Barillot.

Multi-compartment diffusion models (MCM) are increasingly used to characterize the brain white matter microstructure from diffusion MRI. We address the problem of interpolation and averaging of MCM images as a simplification problem based on spectral clustering. As a core part of the framework, we propose novel solutions for the averaging of MCM compartments. Evaluation is performed both on synthetic and clinical data, demonstrating better performance for the "covariance analytic" averaging method. We then present an MCM template of normal controls constructed using the proposed interpolation [23].

7.2.2. The DTI Challenge: Toward Standardized Evaluation of Diffusion Tensor Imaging Tractography for Neurosurgery

Participants: Olivier Commowick, Sylvain Prima.

Diffusion tensor imaging (DTI) tractography reconstruction of white matter pathways can help guide brain tumor resection. However, DTI tracts are complex mathematical objects and the validity of tractography-derived information in clinical settings has yet to be fully established. To address this issue, the DTI Challenge was initiated, an international working group of clinicians and scientists whose goal was to provide standardized evaluation of tractography methods for neurosurgery. The purpose of this empirical study was to evaluate different tractography techniques in the first DTI Challenge workshop. Eight international teams from leading institutions reconstructed the pyramidal tract in four neurosurgical cases presenting with a glioma near the motor cortex. Tractography methods included deterministic, probabilistic, filtered, and global approaches. Standardized evaluation of the tracts consisted in the qualitative review of the pyramidal pathways by a panel of neurosurgeons and DTI experts and the quantitative evaluation of the degree of agreement among methods. The evaluation of tractography reconstructions showed a great inter-algorithm variability. Although most methods found projections of the pyramidal tract from the medial portion of the motor strip, only a few algorithms could trace the lateral projections from the hand, face, and tongue area. In addition, the structure of disagreement among methods was similar across hemispheres despite the anatomical distortions caused by pathological tissues. The DTI Challenge provides a benchmark for the standardized evaluation of tractography methods on neurosurgical data. This study [18] suggests that there are still limitations to the clinical use of tractography for neurosurgical decision making.

7.2.3. Diffusion MRI abnormalities detection with orientation distribution functions: A multiple sclerosis longitudinal study

Participants: Olivier Commowick, Jean-Christophe Ferré, Gilles Edan, Christian Barillot.

We proposed a new algorithm for the voxelwise analysis of orientation distribution functions between one image and a group of reference images [13]. It relies on a generic framework for the comparison of diffusion probabilities on the sphere, sampled from the underlying models. We demonstrated that this method, combined to dimensionality reduction through a principal component analysis, allows for more robust detection of lesions on simulated data when compared to classical tensor-based analysis. We then demonstrated the efficiency of this pipeline on the longitudinal comparison of multiple sclerosis patients at an early stage of the disease: right after their first clinically isolated syndrome (CIS) and three months later. We demonstrated the predictive value of ODF-based scores for the early detection of lesions that will appear or heal.

7.3. EEG and MR Imaging

7.3.1. On the feasibility and specificity of simultaneous EEG and ASL MRI at 3T

Participants: Elise Bannier, Marsel Mano, Isabelle Corouge, Lorraine Perronnet, Christian Barillot.

Brain functional imaging can be performed using several approaches, including EEG, BOLD and ASL MRI. To date, only a few studies have addressed the issue of connecting EEG signal to ASL perfusion. ASL imaging relies on control and label RF pulses, generating alternate gradient patterns as well as higher SAR. The aim of this study was to assess ASL-EEG at 3T in terms of safety as well as EEG and MR signal quality [19].

7.3.2. Symmetrical EEG-FMRI Imaging by Sparse Regularization

Participants: Pierre Maurel, Nicolas Raillard, Saman Noorzadeh, Christian Barillot.

This work [28] considers the problem of brain imaging using simultaneously recorded electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). To this end, we introduce a linear coupling model that links the electrical EEG signal to the hemodynamic response from the blood-oxygen level dependent (BOLD) signal. Both modalities are then symmetrically integrated, to achieve a high resolution in time and space while allowing some robustness against potential decoupling of the BOLD effect. The novelty of the approach consists in expressing the joint imaging problem as a linear inverse problem, which is addressed using sparse regularization. We consider several sparsity-enforcing penalties, which naturally reflect the fact that only few areas of the brain are activated at a certain time, and allow for a fast optimization through proximal algorithms. The significance of the method and the effectiveness of the algorithms are demonstrated through numerical investigations on a spherical head model. This is a joint work with T.Oberlin and R.Gribonval.

7.4. Applications in Neuroradiology and Neurological Disorders

7.4.1. Brain perfusion gender differences using ASL in young adults

Participants: Léa Itmi, Pierre Maurel, Isabelle Corouge, Jean-Christophe Ferré, Christian Barillot.

The use of population models is becoming increasingly important in cerebral imaging, particularly using Arterial Spin Labeling perfusion imaging. Therefore, it is important to know the limits of the models before applying them, to guarantee the reliability of the results. It is now well-known that brain perfusion, in particular cerebral blood flow (CBF), changes with age, and this effect needs to be taken into account when evaluating brain perfusion images. But gender differences have not been well studied yet. It is known that female brain perfusion is, in average, higher than male brain perfusion, but only few studies have investigated whether some regional perfusion differences exist or not. This work aims to assess whether, as for the age, gender differences should be taken into account when analyzing brain perfusion images. We then focus on adult subjects and study the CBF gender differences. We compared the raw CBF means and the means after normalization, we also investigated perfusion asymetries. We used atlases for the region comparisons and the General Linear Model for the voxel level. Our results confirmed that women have a higher CBF than men, and showed that this difference can be suppressed with a normalization process, but no specific major regional difference or asymmetry was found.

7.4.2. Arterial Spin Labeling Motor Activation Presurgical Mapping for Brain Tumor Resection

Participants: Isabelle Corouge, Elise Bannier, Jean-Christophe Ferré.

Functional Arterial Spin Labeling (fASL) has demonstrated its greater specificity as a marker of neuronal activity than the reference BOLD fMRI for motor activation mapping in healthy volunteers. Motor fASL is yet to be investigated in the context of tumors, under the assumption that fASL would be less sensitive to venous contamination induced by the hemodynamics remodeling in the tumor vicinity than BOLD fMRI. As the arterial transit time may be shortened in activation areas, this preliminary study explores the ability of fASL to map the motor areas at different post-labeling delays (PLD) in healthy subjects and patient with brain tumor [21].

7.4.3. Dynamic assessment of macrophages infiltration and tissue damage in MS lesions

Participants: Anne Kerbrat, Benoit Combès, Olivier Commowick, Jean-Christophe Ferré, Elise Bannier, Christian Barillot, Gilles Edan.

Inflammation is a dynamic and complex process that could be beneficial when it supports tissue repair but also detrimental when excessive, leading to worsen tissue injury. In multiple sclerosis, it is well known from pathological and MRI studies that the prognostic between white matter lesions differed at the lesion level. Thus, 10 to 30% of T2 hyperintense lesions are seen as area of persistent hypointensity on T1-w images. These T1 hypointensity are areas of pathologically confirmed severe axonal loss. Complementary, quantitative MRI such as Diffusion imaging, magnetization transfer imaging and relaxometry can quantify and characterize

tissue changes on MRI before, during, and after the evolution of a new MRI-detected lesion. They are related to damage to myelin and axons. However, identifying in vivo the dynamic pathophysiological processes that leads to these various degree of demyelination and axonal loss in MS lesions remained challenging. In recent year, molecular and cellular imaging of the inflammatory process have been developed. Although some techniques remains at the pre-clinical level, MRI using non targeted USPIO as contrast agent can be used in MS patients. USPIO are phagocyted in periphery by macrophages and migrate to the central nervous system to characterize in vivo macrophages infiltrations within lesions. The association of cellular imaging and longitudinal quantitative MRI consist of a great opportunity to assess more specifically the overall process. In a recent study from our group, we demonstrated that infiltration of activated macrophages evidenced by USPIO enhancement, was present at the onset of MS and associated with higher local loss of tissue structure [17]. This year, we pursued this work by analyzing a longitudinal study with USPIO infusion every 3 months, associated with quantitative MRI assessment including MTI, diffusion imaging and relaxometry with the objectives of describing relationships between macrophages infiltration and quantitative MRI metrics reflecting tissue structure along time.

7.4.4. The effect of water suppression on the hepatic lipid quantification, as assessed by the LCModel, in a preclinical and clinical scenario

Participant: Elise Bannier.

This work investigates the effect of water suppression on the hepatic lipid quantification, using the LCModel. MR spectra with and without water suppression were acquired in the liver of mice at 4.7 T and patients at 3 T, and processed with the LCModel. The Cramer-Rao Lower Bound (CRLB) values of the seven lipid resonances were determined to assess the impact of water suppression on hepatic lipid quantification. A paired t test was used for comparison between the CRLBs obtained with and without water suppression. For the preclinical data, in the high (low) fat fraction subset an overall impairment in hepatic lipid quantification, i.e. an increase of CRLBs (no significant change of CRLBs) was observed in spectra acquired with water suppression. For the clinical data, there were no substantial changes in the CRLB with water suppression. Because (1) the water suppression does not overall improve the quantification of the lipid resonances and (2) the MR spectrum without water suppression is always acquired for fat fraction calculation, the optimal data-acquisition strategy for liver MRS is to acquire only the MR spectrum without water suppression. For quantification of hepatic lipid resonances, it is advantageous to perform MR spectroscopy without water suppression in a clinical and preclinical scenario (at moderate fields) [14].

7.5. Management of Information in Neuroimaging

Participants: Michael Kain, Olivier Commowick, Elise Bannier, Inès Fakhfakh, Justine Guillaumont, Florent Leray, Yao Yao, Christian Barillot.

The major topic that is addressed in this period concern the sharing of data and processing tools in neuroimaging (through the “Programme d’Investissement d’Avenir” project such as OFSEP and FLI-IAM) which led to build a suitable architecture to share images and processing tools, started from the NeuroBase project (supported by the French Ministry of Research). Our overall goal within these projects is to set up a computer infrastructure to facilitate the sharing of neuroimaging data, as well as image processing tools, in a distributed and heterogeneous environment. These consortium gathered expertise coming from several complementary domains of expertise: image processing in neuroimaging, workflows and GRID computing, ontology development and ontology-based mediation. This enables a large variety of users to diffuse, exchange or reach neuroimaging information with appropriate access means, in order to be able to retrieve information almost as easily as if the data were stored locally by means of the "cloud computing" Storage as a Service (SaaS) concept As an example, the Shanoir environment has been succesfully deployed to the Neurinfo platform were it is routinely used to manage images of the research studies. It is also currently being deployed for two large projets: OFSEP ("Observatoire Français de la Sclérose en Plaques") where up to 30000 patients will be acquired on a ten years frame, and the Image Analysis and Mangment (IAM) node of the France Life Imaging national infrastructure (FLI-IAM). Our team fulfills multiple roles in this nation-wide FLI project. Christian Barillot

is the chair of the IAM node, Olivier Commowick is participating in the working group workflow and image processing and Michael Kain is the technical manager of the node. Apart from the team members, software solutions like medInria and Shanoir are part of the final software platform.

ASAP Project-Team

6. New Results

6.1. Models and Theory of Distributed Systems

6.1.1. Asynchronous Byzantine Systems: From Multivalued to Binary Consensus with t < n/3, O(n^2) Messages, O(1) Time, and no Signature

Participant: Michel Raynal.

This work [39] presents a new algorithm that reduces multivalued consensus to binary consensus in an asynchronous message-passing system made up of n processes where up to t may commit Byzantine failures. This algorithm has the following noteworthy properties: it assumes t < n/3 (and is consequently optimal from a resilience point of view), uses O(n^2) messages, has a constant time complexity, and does not use signatures. The design of this reduction algorithm relies on two new all-to-all communication abstractions. The first one allows the non-faulty processes to reduce the number of proposed values to c, where c is a small constant. The second communication abstraction allows each non-faulty process to compute a set of (proposed) values such that, if the set of a non-faulty process contains a single value, then this value belongs to the set of any non-faulty process. Both communication abstractions have an O(n^2) message complexity and a constant time complexity. The reduction of multivalued Byzantine consensus to binary Byzantine consensus is then a simple sequential use of these communication abstractions. To the best of our knowledge, this is the first asynchronous message-passing algorithm that reduces multivalued consensus to binary consensus with O(n^2) messages and constant time complexity (measured with the longest causal chain of messages) in the presence of up to t < n/3 Byzantine processes, and without using cryptography techniques. Moreover, this reduction algorithm tolerates message reordering by Byzantine processes.

This work was done in collaboration with Achour Mostefaoui from the LINA laboratory in Nantes.

6.1.2. Atomic Read/Write Memory in Signature-free Byzantine Asynchronous Message-passing Systems

Participant: Michel Raynal.

In this work [54] we designed a signature-free distributed algorithm which builds an atomic read/write shared memory on top of an n-process asynchronous message-passing system in which up to t < n/3 processes may commit Byzantine failures. From a conceptual point of view, this algorithm is designed to be as close as possible to the algorithm proposed by Attiya, Bar-Noy and Dolev (JACM 1995), which builds an atomic register in an n-process asynchronous message-passing system where up to t < n/2 processes may crash. The proposed algorithm is particularly simple. It does not use cryptography to cope with Byzantine processes, and is optimal from a t-resilience point of view (t < n/3). A read operation requires O(n) messages, and a write operation requires O(n 2) messages.

This work was done in collaboration with Achour Mostefaoui, Matoula Petrolia, and Claude Jard from the LINA laboratory in Nantes.

6.1.3. Intrusion-Tolerant Broadcast and Agreement Abstractions in the Presence of Byzantine Processes

Participant: Michel Raynal.

A process commits a Byzantine failure when its behavior does not comply with the algorithm it is assumed to execute. Considering asynchronous message-passing systems, this work [18] presents distributed abstractions, and associated algorithms, that allow non-faulty processes to correctly cooperate, despite the uncertainty created by the net effect of asynchrony and Byzantine failures. These abstractions are broadcast abstractions (namely, no-duplicity broadcast, reliable broadcast, and validated broadcast), and agreement abstraction (namely, consensus). While no-duplicity broadcast and reliable broadcast are well-known one-to-all communication abstractions, validated broadcast is a new all-to-all communication abstraction designed to address agreement problems. After having introduced these abstractions, we also presented an algorithm implementing validated broadcast on top of reliable broadcast. Then we presented two consensus algorithms, which are reductions of multivalued consensus to binary consensus. The first one is a generic algorithm, that can be instantiated with unreliable broadcast or no-duplicity broadcast, while the second is a consensus algorithm based on validated broadcast. Finally, a third algorithm is presented that solves the binary consensus problem. This algorithm is a randomized algorithm based on validated broadcast and a common coin. The presentation of all the abstractions and their algorithms is done incrementally. This work was done in collaboration with Achour Mostefaoui from the LINA laboratory in Nantes.

6.1.4. Minimal Synchrony for Asynchronous Byzantine Consensus

Participants: Michel Raynal, Zohir Bouzid.

Solving the consensus problem requires in one way or another that the underlying system satisfies some synchrony assumption. Considering an asynchronous message-passing system of n processes where (a) up to t < n/3 may commit Byzantine failures, and (b) each pair of processes is connected by two uni-directional channels (with possibly different timing properties), this work [24] investigates the synchrony assumption required to solve consensus, and presents a signature-free consensus algorithm whose synchrony requirement is the existence of a process that is an eventual t+1bisource. Such a process p is a correct process that eventually has (a) timely input channels from t correct processes and (b) timely output channels to t correct processes (these input and output channels can connect p to different subsets of processes). As this synchrony condition was shown to be necessary and sufficient in the stronger asynchronous system model (a) enriched with message authentication, and (b) where the channels are bidirectional and have the same timing properties in both directions, it follows that it is also necessary and sufficient in the weaker system model considered in this work. In addition to the fact that it closes a long-lasting problem related to Byzantine agreement, a noteworthy feature of the proposed algorithm lies in its design simplicity, which is a first-class property.

This work was done in collaboration with Achour Mostefaoui from the LINA laboratory in Nantes.

6.1.5. Signature-Free Asynchronous Binary Byzantine Consensus with t<n/3, O(n^2) Messages, and O(1) Expected Time

Participant: Michel Raynal.

This work [17] is on broadcast and agreement in asynchronous message-passing systems made up of n processes, and where up to t processes may have a Byzantine Behavior. Its first contribution is a powerful, yet simple, all-to-all broadcast communication abstraction suited to binary values. This abstraction, which copes with up to t < n/3 Byzantine processes, allows each process to broadcast a binary value, and obtain a set of values such that (1) no value broadcast only by Byzantine processes can belong to the set of a correct process, and (2) if the set obtained by a correct process contains a single value v, then the set obtained by any correct process contains v. The second contribution of this work is a new round-based asynchronous consensus algorithm that copes with up to t < n/3 Byzantine processes. This algorithm is based on the previous binary broadcast abstraction and a weak common coin. In addition of being signature-free and optimal with respect to the value of t, this consensus algorithm has several noteworthy properties: the expected number of rounds to decide is constant; each round is composed of a constant number of communication steps and involves O(n^2) messages; each message is composed of a round number plus a constant number of bits. Moreover, the algorithm tolerates message reordering by the adversary (i.e., the Byzantine processes). This work was done in collaboration with Achour Mostefaoui from the LINA laboratory in Nantes, and Hamouma Moumen from Université de Béjaïa.

6.1.6. Specifying Concurrent Problems: Beyond Linearizability and up to Tasks

Participants: Michel Raynal, Zohir Bouzid.

Tasks and objects are two predominant ways of specifying distributed problems. A task specifies for each set of processes (which may run concurrently) the valid outputs of the processes. An object specifies the outputs the object may produce when it is accessed sequentially. Each one requires its own implementation notion, to tell when an execution satisfies the specification. For objects linearizability is commonly used, while for tasks implementation notions are less explored. Sequential specifications are very convenient, especially important is the locality property of linearizability, which states that linearizable objects compose for free into a linearizable object. However, most well-known tasks have no sequential specification. Also, tasks have no clear locality property. This work [26] introduces the notion of interval-sequential object. The corresponding implementation notion of interval-linearizability generalizes linearizability. Interval-linearizability allows to specify any task. However, there are sequential one-shot objects that cannot be expressed as tasks, under the simplest interpretation of a task. We also showed that a natural extension of the notion of a task is expressive enough to specify any interval-sequential object.

This work was done in collaboration with Armando Castaneda and Sergio Rajsbaumm from UNAM, Mexico.

6.1.7. Test-and-Set in Optimal Space

Participant: George Giakkoupis.

The test-and-set object is a fundamental synchronization primitive for shared memory systems. In [35] we address the number of registers (supporting atomic reads and writes) required to implement a one-shot test-and-set object in the standard asynchronous shared memory model with n processes. The best lower bound is $\log n - 1$ for obstruction-free and deadlock-free implementations, and recently a deterministic obstruction-free implementation using $O(\sqrt{n})$ registers was presented.

In [35] we close the gap between these existing upper and lower bounds by presenting a deterministic obstruction-free implementation of a one-shot test-and-set object from $\Theta(\log n)$ registers of size $\Theta(\log n)$ bits. Combining our obstruction-free algorithm with techniques from previous research, we also obtain a randomized wait-free test-and-set algorithm from $\Theta(\log n)$ registers, with expected step-complexity $\Theta(\log^* n)$ against the oblivious adversary. The core tool in our algorithm is the implementation of a deterministic obstruction-free *sifter* object, using only 6 registers. If k processes access a sifter, then when they have terminated, at least one and at most $\mathsf{floor}(2k+1)/3$ processes return "win" and all others return "lose".

This is a joint work with Maryam Helmi (U. of Calgary), Lisa Higham (U. of Calgary), and Philipp Woelfel (U. of Calgary), supported by the RADCON Inria Associate Team.

6.2. Graph and Probabilistic Algorithms

6.2.1. On the Quadratic Shortest Path Problem

Participant: Davide Frey.

Finding the shortest path in a directed graph is one of the most important combinatorial optimization problems, having applications in a wide range of fields. In its basic version, however, the problem fails to represent situations in which the value of the objective function is determined not only by the choice of each single arc, but also by the combined presence of pairs of arcs in the solution. In this work [40] we model these situations as a Quadratic Shortest Path Problem, which calls for the minimization of a quadratic objective function subject to shortest-path constraints. We prove strong NP-hardness of the problem and analyze polynomially solvable special cases, obtained by restricting the distance of arc pairs in the graph that appear jointly in a quadratic monomial of the objective function. Based on this special case and problem structure, we devise fast lower bounding procedures for the general problem and show computationally that they clearly outperform other approaches proposed in the literature in terms of its strength.

6.2.2. Tight Bounds on Vertex Connectivity Under Vertex Sampling

Participant: George Giakkoupis.

A fundamental result by Karger (SODA 1994) states that for any λ-edge-connected graph with n nodes, independently sampling each edge with probability $p = \Omega(\log n/\lambda)$ results in a graph that has edge connectivity $\Omega(\lambda p)$, with high probability. In [27] we prove the analogous result for vertex connectivity, when sampling vertices. We show that for any k-vertex-connected graph G with n nodes, if each node is independently sampled with probability $p = \Omega(\sqrt{\log n/k})$, then the subgraph induced by the sampled nodes has vertex connectivity $\Omega(kp^2)$, with high probability. This bound improves upon the recent results of Censor-Hillel et al. (SODA 2014) and is existentially optimal.

This is a joint work with Keren Censor-Hillel (Technion), Mohsen Ghaffari (MIT), Bernhard Haeupler (Carnegie Mellon U.), and Fabian Kuhn (U. of Freiburg).

6.3. Scalable Systems

6.3.1. Being prepared in a sparse world: the case of KNN graph construction

Participants: Anne-Marie Kermarrec, Nupur Mittal, Francois Taiani.

This work presents KIFF [41], a generic, fast and scalable KNN graph construction algorithm. KIFF directly exploits the bipartite nature of most datasets to which KNN algorithms are applied. This simple but powerful strategy drastically limits the computational cost required to rapidly converge to an accurate KNN solution, especially for sparse datasets. Our evaluation on a representative range of datasets show that KIFF provides, on average, a speed-up factor of 14 against recent state-of-the art solutions while improving the quality of the KNN approximation by 18

This work was done in collaboration with Antoine Boutet from CNRS, Laboratoire Hubert Curien, Saint-Etienne, France.

6.3.2. Cheap and Cheerful: Trading Speed and Quality for Scalable Social Recommenders

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

Recommending appropriate content and users is a critical feature of on-line social networks. Computing accurate recommendations on very large datasets can however be particularly costly in terms of resources, even on modern parallel and distributed infrastructures. As a result, modern recommenders must generally trade-off quality and computational cost to reach a practical solution. This trade-off has however so far been largely left unexplored by the research community, making it difficult for practitioners to reach informed design decisions. In this work [37], we investigate to which extent the additional computing costs of advanced recommendation techniques based on supervised classifiers can be balanced by the gains they bring in terms of quality. In particular, we compare these recommenders against their unsupervised counterparts, which offer lightweight and highly scalable alternatives. We propose a thorough evaluation comparing 11 classifiers against 7 lightweight recommenders on a real Twitter dataset. Additionally, we explore data grouping as a method to reduce computational costs in a distributed setting while improving recommendation quality. We demonstrate how classifiers trained using data grouping can reduce their computing time by 6 while improving recommendations up to 22% when compared with lightweight solutions.

6.3.3. Fast Nearest Neighbor Search

Participants: Fabien André, Anne-Marie Kermarrec.

Nearest Neighbor (NN) search in high dimension is an important feature in many applications, such as multimedia databases, information retrieval or machine learning. Product Quantization (PQ) is a widely used solution which offers high performance, i.e., low response time while preserving a high accuracy. PQ represents high-dimensional vectors by compact codes. Large databases can therefore be stored in memory, allowing NN queries without resorting to slow I/O operations. PQ computes distances to neighbors using cache-resident lookup tables, thus its performance remains limited by (i) the many cache accesses that the algorithm requires, and (ii) its inability to leverage SIMD instructions available on modern CPUs.

To address these limitations, we designed a novel algorithm, PQ Fast Scan [19], that transforms the cache-resident lookup tables into small tables, sized to fit SIMD registers. This transformation allows (i) in-register lookups in place of cache accesses and (ii) an efficient SIMD implementation. PQ Fast Scan has the exact same accuracy as PQ, while having 4 to 6 times lower response time (e.g., for 25 million vectors, scan time is reduced from 74ms to 13ms).

This work was done in collaboration with Nicolas Le Scouarnec.

6.3.4. Holons: towards a systematic approach to composing systems of systems

Participants: Yérom-David Bromberg, François Taïani.

The world's computing infrastructure is increasingly differentiating into self-contained distributed systems with various purposes and capabilities (e.g. IoT installations, clouds, VANETs, WSNs, CDNs,. . .). Furthermore, such systems are increasingly being composed to generate systems of systems that offer value-added functionality. Today, however, system of systems composition is typically ad-hoc and fragile. It requires developers to possess an intimate knowledge of system internals and low-level interactions between their components. In this work [21], we outline a vision and set up a research agenda towards the generalised programmatic construction of distributed systems as compositions of other distributed systems. Our vision, in which we refer uniformly to systems and to compositions of systems as holons, employs code generation techniques and uses common abstractions, operations and mechanisms at all system levels to support uniform system of systems composition. We believe our holon approach could facilitate a step change in the convenience and correctness with which systems of systems can be built, and open unprecedented opportunities for the emergence of new and previously-unenvisaged distributed system deployments, analogous perhaps to the impact the mashup culture has had on the way we now build web applications.

This work was done in collaboration with Gordon Blair Geoff Coulson, and Yehia Elkhatib from Lancaster University (UK), Laurent Réveillère from University of Bordeaux / Labri, and Heverson Borba Ribeiro and Etienne Rivière from University of Neuchâtel (Switzerland).

6.3.5. Hybrid datacenter scheduling

Participant: Anne-Marie Kermarrec.

We address the problem of efficient scheduling of large clusters under high load and heterogeneous workloads. A heterogeneous workload typically consists of many short jobs and a small number of large jobs that consume the bulk of the cluster's resources.

Recent work advocates distributed scheduling to overcome the limitations of centralized schedulers for large clusters with many competing jobs. Such distributed schedulers are inherently scalable, but may make poor scheduling decisions because of limited visibility into the overall resource usage in the cluster. In particular, we demonstrate that under high load, short jobs can fare poorly with such a distributed scheduler.

We propose instead a new hybrid centralized/ distributed scheduler, called Hawk. In Hawk, long jobs are scheduled using a centralized scheduler, while short ones are scheduled in a fully distributed way. Moreover, a small portion of the cluster is reserved for the use of short jobs. In order to compensate for the occasional poor decisions made by the distributed scheduler, we propose a novel and efficient randomized work-stealing algorithm.

We evaluate Hawk using a trace-driven simulation and a prototype implementation in Spark. In particular, using a Google trace, we show that under high load, compared to the purely distributed Sparrow scheduler, Hawk improves the 50th and 90th percentile runtimes by 80% and 90% for short jobs and by 35% and 10% for long jobs, respectively. Measurements of a prototype implementation using Spark on a 100-node cluster confirm the results of the simulation. This work has been done in the context of the Inria/epfl research center and in collaboration with Pamela delgado, Florin Dinu and Willy Zwaenepoel from EPFL and published in Usenix ATC in 2015 [30].

6.3.6. Out-of-core KNN Computation

Participants: Nitin Chiluka, Anne-Marie Kermarrec, Javier Olivares.

This work proposes a novel multi threading approach to compute KNN on large datasets by leveraging both disk and main memory efficiently. The main rationale of our approach is to minimize random accesses to disk, maximize sequential access to data and efficient usage of only a fraction of the available memory. This approach is evaluated by comparing its performance with a fully in-memory implementation of KNN, in terms of execution time and memory consumption. This multithreading approach outperforms the in-memory baseline in all cases when the large dataset does not fit in memory.

6.3.7. Scaling Out Link Prediction with SNAPLE

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

A growing number of organizations are seeking to analyze extra large graphs in a timely and resource-efficient manner. With some graphs containing well over a billion elements, these organizations are turning to distributed graph-computing platforms that can scale out easily in existing data-centers and clouds. Unfortunately such platforms usually impose programming models that can be ill suited to typical graph computations, fundamentally undermining their potential benefits. In this work [38], we consider how the emblematic problem of link-prediction can be implemented efficiently in gather-apply-scatter (GAS) platforms, a popular distributed graph-computation model. Our proposal, called Snaple, exploits a novel highly-localized vertex scoring technique, and minimizes the cost of data flow while maintaining prediction quality. When used within GraphLab, Snaple can scale to very large graphs that a standard implementation of link prediction on GraphLab cannot handle. More precisely, we show that Snaple can process a graph containing 1.4 billions edges on a 256 cores cluster in less than three minutes, with no penalty in the quality of predictions. This result corresponds to an over-linear speedup of 30 against a 20-core standalone machine running a non-distributed state-of-the-art solution.

6.3.8. Similitude: Decentralised Adaptation in Large-Scale P2P Recommenders

Participants: Davide Frey, Anne-Marie Kermarrec, Pierre-Louis Roman, François Taïani.

Decentralised recommenders have been proposed to deliver privacy-preserving, personalised and highly scalable on-line recommendations. Current implementations tend, however, to rely on a hard-wired similarity metric that cannot adapt. This constitutes a strong limitation in the face of evolving needs. In this work [33], we propose a framework to develop dynamically adaptive decentralized recommendation systems. Our proposal supports a decentralised form of adaptation, in which individual nodes can independently select, and update their own recommendation algorithm, while still collectively contributing to the overall system's mission.

This work was done in collaboration with Christopher Maddock and Andreas Mauthe (Univ. of Lancaster, UK).

6.3.9. Transactional Memory Recommenders

Participant: Anne-Marie Kermarrec.

The Transactional Memory (TM) paradigm promises to greatly simplify the development of concurrent applications. This led, over the years, to the creation of a plethora of TM implementations delivering wide ranges of performance across workloads. Yet, no universal TM implementation fits each and every workload. In fact, the best TM in a given workload can reveal to be disastrous for another one. This forces developers to face the complex task of tuning TM implementations, which significantly hampers the wide adoption of TMs. In this work, we address the challenge of automatically identifying the best TM implementation for a given workload. Our proposed system, ProteusTM, hides behind the TM interface a large library of implementations. Under the hood, it leverages an innovative, multi-dimensional online optimization scheme, combining two popular machine learning techniques: Collaborative Filtering and Bayesian Optimization. We integrated ProteusTM in GCC and demonstrated its ability to switch TM implementations and adapt several configuration parameters (e.g., number of threads). We extensively evaluated ProteusTM, obtaining average performance 3% less than the optimal, and gains up to 100 over static alternatives.

This work has been done in collaboration with Rachid Guerraoui from EPFL, Diego Didona Nuno Diegues, Ricardo Neves and Paolo Romano from INESC, Lisboa) and will be published in ASPLOS 2016 [31].

6.3.10. Want to scale in centralized systems? Think P2P

Participants: Anne-Marie Kermarrec, François Taïani.

Peer-to-peer (P2P) systems have been widely researched over the past decade, leading to highly scalable implementations for a wide range of distributed services and applications. A P2P system assigns symmetric roles to machines, which can act both as client and server. This distribution of responsibility alleviates the need for any central component to maintain a global knowledge of the system. Instead, each peer takes individual decisions based on a local and limited knowledge of the rest of the system, providing scalability by design. While P2P systems have been successfully applied to a wide range of distributed applications (multicast, routing, caches, storage, pub-sub, video streaming), with some highly visible successes (Skype, Bitcoin), they tend to have fallen out of fashion in favor of a much more cloud-centric vision of the current Internet. We think this is paradoxical, as cloud-based systems are themselves large-scale, highly distributed infrastructures. They reside within massive, densely interconnected datacenters, and must execute efficiently on an increasing number of machines, while dealing with growing volumes of data. Today even more than a decade ago, large-scale systems require scalable designs to deliver efficient services. In this work [16] we argue that the local nature of P2P systems is key for scalability regardless whether a system is eventually deployed on a single multi-core machine, distributed within a data center, or fully decentralized across multiple autonomous hosts. Our claim is backed by the observation that some of the most scalable services in use today have been heavily influenced by abstractions and rationales introduced in the context of P2P systems. Looking to the future, we argue that future large-scale systems could greatly benefit from fully decentralized strategies inspired from P2P systems. We illustrate the P2P legacy through several examples related to Cloud Computing and Big Data, and provide general guidelines to design large-scale systems according to a P2P philosophy.

6.3.11. WebGC: Browser-based gossiping

Participants: Raziel Carvajal Gomez, Davide Frey, Anne-Marie Kermarrec.

The advent of browser-to-browser communication technologies like WebRTC has renewed interest in the peer-to-peer communication model. However, the available WebRTC code base still lacks important components at the basis of several peer-to-peer solutions. Through a collaboration with Mathieu Simonin from the Inria SED in the context of the Brow2Brow ADT project, we started to tackle this problem by proposing WebGC, a library for gossip-based communication between web browsers. Due to their inherent scalability, gossip-based, or epidemic protocols constitute a key component of a large number of decentralized applications. WebGC thus represents an important step towards their wider spread. We demonstrated the final version of the library at WISE 2015 [53].

6.4. Privacy in User Centric Applications

6.4.1. Collaborative Filtering Under a Sybil Attack: Similarity Metrics do Matter!

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

Whether we are shopping for an interesting book or selecting a movie to watch, the chances are that a recommendation system will help us decide what we want. Recommendation systems collect information about our own preferences, compare them to those of other users, and provide us with suggestions on a variety of topics. But is the information gathered by a recommendation system safe from potential attackers, be them other users, or companies that access the recommendation system? And above all, can service providers protect this information while still providing effective recommendations? In this work, we analyze the effect of Sybil attacks on collaborative-filtering recommendation systems, and discuss the impact of different similarity metrics in the trade-off between recommendation quality and privacy. Our results, on a state-of-the-art recommendation framework and on real datasets show that existing similarity metrics exhibit a wide range of behaviors in the presence of Sybil attacks. Yet, they are all subject to the same trade off: Sybil resilience for recommendation quality. We therefore propose and evaluate a novel similarity metric that combines the best of both worlds: a low RMSE score with a prediction accuracy for Sybil users of only a few percent. A preliminary version of this work was published at EuroSec 2015 [32].

This work was done in collaboration with Antoine Boutet, and Rachid Guerraoui.

6.4.2. Decentralized view prediction for global content placement

Participants: Stéphane Delbruel, Davide Frey, François Taïani.

A large portion of today's Internet traffic originates from streaming and video services. Storing, indexing, and serving these videos is a daily engineering challenge that requires increasing amounts of efforts and infrastructures. One promising direction to improve video services consists in predicting at upload time where and when a new video might be viewed, thereby optimizing placement and caching decisions. Implementing such a prediction service in a scalable manner poses significant technical challenges. In this work [28], we address these challenges in the context of a decentralized storage system consisting of set-top boxes or end nodes. Specifically, we propose a novel data placement algorithm that exploits information about the tags associated with existing content, such as videos, and uses it to infer the number of views that newly uploaded content will have in each country.

6.4.3. Distance-Based Differential Privacy in Recommenders

Participant: Anne-Marie Kermarrec.

The upsurge in the number of web users over the last two decades has resulted in a significant growth of online information. This information growth calls for recommenders that personalize the information proposed to each individual user. Nevertheless, personalization also opens major privacy concerns. We designed D2P, a novel protocol that ensures a strong form of differential privacy, which we call distance-based differential privacy, and which is particularly well suited to recommenders. D2P avoids revealing exact user profiles by creating altered profiles where each item is replaced with another one at some distance. We evaluate D2P analytically and experimentally on MovieLens and Jester datasets and compare it with other private and non-private recommenders. This work has been done in the context of the Web-Alter-ego Google focused award and in collaboration with Rachid guerraoui, Rhicheek Patra and Masha Taziki from EPFL and published in PVLVB in 2015 [15].

6.4.4. Privacy-Conscious Information Diffusion in Social Networks

Participants: George Giakkoupis, Arnaud Jégou, Anne-Marie Kermarrec, Nupur Mittal.

This work presents a distributed algorithm – Riposte [47], for information dissemination in social networks which guarantees to preserve the privacy of its users. RIPOSTE ensures that information spreads widely if and only if a large fraction of users find it interesting, and this is done in a "privacy-conscious" manner, namely without revealing the opinion of any individual user. Whenever an information item is received by a user, RIPOSTE decides to either forward the item to all the user's neighbors, or not to forward it to anyone. The decision is randomized and is based on the user's (private) opinion on the item, as well as on an upper bound s on the number of user's neighbors that have not received the item yet.

This work was done in collaboration with Rachid Guerraoui from EPFL, Switzerland.

6.4.5. Hide & Share: Landmark-based Similarity for Private knn Computation

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

Computing k-nearest-neighbor graphs constitutes a fundamental operation in a variety of data-mining applications. As a prominent example, user-based collaborative-filtering provides recommendations by identifying the items appreciated by the closest neighbors of a target user. As this kind of applications evolve, they will require KNN algorithms to operate on more and more sensitive data. This has prompted researchers to propose decentralized peer-to-peer KNN solutions that avoid concentrating all information in the hands of one central organization. Unfortunately, such decentralized solutions remain vulnerable to malicious peers that attempt to collect and exploit information on participating users.

In this work [22], we seek to overcome this limitation by proposing H&S (Hide & Share), a novel landmark-based similarity mechanism for decentralized KNN computation. Landmarks allow users (and the associated peers) to estimate how close they lay to one another without disclosing their individual profiles.

We evaluate H&S in the context of a user-based collaborative-filtering recommender with publicly available traces from existing recommendation systems. We show that although landmark-based similarity does disturb similarity values (to ensure privacy), the quality of the recommendations is not as significantly hampered. We also show that the mere fact of disturbing similarity values turns out to be an asset because it prevents a malicious user from performing a profile reconstruction attack against other users, thus reinforcing users' privacy. Finally, we provide a formal privacy guarantee by computing the expected amount of information revealed by H&S about a user's profile.

This work was done in collaboration with Antoine Boutet from the University of St. Etienne, and with Jingjing Wang and Rachid Guerraoui from EPFL, Switzerland.

ASCOLA Project-Team

6. New Results

6.1. Highlights of the year

Nicolas Tabareau has been awarded a starting grant from the European Research Council (ERC), the most prestigious type of research projects of the European Union for young researchers. From 2015–2020 he will pursue research on "CoqHoTT: Coq for Homotopy Type Theory."

In the domain of resource management notably for Cloud infrastructures, the team has produced several very visible results. These include contributions to popular and new simulation tools and platforms [17], [27], [28] as well as new techniques for the energy-efficient execution of Cloud applications [15].

On the topics of software composition and programming languages, the team has, among others, two remarkable results: a new notion of effect capabilities and corresponding monadic analysis techniques [14] as well as the first comprehensive survey of domain-specific aspect languages [13].

6.2. Programming Languages

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

6.2.1. Formal Methods, logics and type theory

This year we have proposed "Gradual Certified Programming" as a bridge between type-based expressive proofs and programming languages, have extended previous type theories by new homotopy-based means, and have introduced "effect capabilities" to control monad-based effects in Haskell.

6.2.1.1. Gradual Certified Programming in Coq

Expressive static typing disciplines are a powerful way to achieve high-quality software. However, the adoption cost of such techniques should not be under-estimated. Just like gradual typing allows for a smooth transition from dynamically-typed to statically-typed programs, it seems desirable to support a gradual path to certified programming. We have explored gradual certified programming in Coq [33], providing the possibility to postpone the proofs of selected properties, and to check "at runtime" whether the properties actually hold. Casts can be integrated with the implicit coercion mechanism of Coq to support implicit cast insertion à la gradual typing. Additionally, when extracting Coq functions to mainstream languages, our encoding of casts supports lifting assumed properties into runtime checks. Much to our surprise, it is not necessary to extend Coq in any way to support gradual certified programming. A simple mix of type classes and axioms makes it possible to bring gradual certified programming to Coq in a straightforward manner.

6.2.1.2. Homotopy Hypothesis in Type Theory

In classical homotopy theory, the homotopy hypothesis asserts that the fundamental omega-groupoid construction induces an equivalence between topological spaces and weak omega-groupoids. In the light of Voevodsky's univalent foundations program, which puts forward an interpretation of types as topological spaces, we have considered the question of transposing the homotopy hypothesis to type theory [16]. Indeed such a transposition could stand as a new approach to specifying higher inductive types. Since the formalization of general weak omega-groupoids in type theory is a difficult task, we have only taken a first step towards this goal, which consists in exploring a shortcut through strict omega-categories. The first outcome is a satisfactory type-theoretic notion of strict omega-category, which has hsets of cells in all dimensions. For this notion, defining the 'fundamental strict omega-category' of a type seems out of reach. The second outcome is an 'incoherently strict' notion of type-theoretic omega-category, which admits arbitrary types of cells in all dimensions. These are the 'wild' omega-categories of the title. They allow the definition of a 'fundamental wild omega-category' map, which leads to our (partial) homotopy hypothesis for type theory (stating an adjunction, not an equivalence). All of our results have been formalized in the Coq proof assistant. Our formalization makes systematic use of the machinery of coinductive types.

6.2.1.3. Effect Capabilities For Haskell

Computational effects complicate the tasks of reasoning about and maintaining software, due to the many kinds of interferences that can occur. While different proposals have been formulated to alleviate the fragility and burden of dealing with specific effects, such as state or exceptions, there is no prevalent robust mechanism that addresses the general interference issue. Building upon the idea of capability-based security, we have proposed effect capabilities [14] as an effective and flexible manner to control monadic effects and their interferences. Capabilities can be selectively shared between modules to establish secure effect-centric coordination. We have further refined capabilities with type-based permission lattices to allow fine-grained decomposition of authority. An implementation of effect capabilities in Haskell has been done, using type classes to establish a way to statically share capabilities between modules, as well as to check proper access permissions to effects at compile time.

6.2.1.4. Correct Refactoring Tools

Most integrated development environments provide refactoring tools. However, these tools are often unreliable. As a consequence, developers have to test their code after applying an automatic refactoring.

Refactoring tools for industrial languages are difficult to test and verify. We have developed a refactoring operation for C programs (renaming of global variables) for which we have proved that it preserves the set of possible behaviors of the transformed programs [39]. That proof of correctness relies on the operational semantics provided by CompCert C in Coq. We have also proved some properties of the transformation which are used to establish properties of a composed refactoring operations.

6.2.2. Language Mechanisms

This year we have contributed new results on domain-specific aspect languages, concurrent event-based programming, model transformations as well as the relationship between functional and constraint programming.

Furthermore, we have proposed language support for the definition and enforcement of security properties, in particular related to the accountability of service-based systems, see Sec. 6.3 .

6.2.2.1. Domain-Specific Aspect Languages

Domain-Specific Aspect Languages (DSALs) are Domain-Specific Languages (DSLs) designed to express crosscutting concerns. Compared to DSLs, their aspectual nature greatly amplifies the language design space. In the context of the Associate Team RAPIDS/REAL, we have structured this space in order to shed light on and compare the different domain-specific approaches to deal with crosscutting concerns [13]. We have reported on a corpus of 36 DSALs covering the space, discussed a set of design considerations and provided a taxonomy of DSAL implementation approaches. This work serves as a frame of reference to DSAL and DSL researchers, enabling further advances in the field, and to developers as a guide for DSAL implementations.

6.2.2.2. Concurrent Event-Based Programming

The advanced concurrency abstractions provided by the Join calculus overcome the drawbacks of low-level concurrent programming techniques. However, with current approaches, the coordination logic involved in complex coordination schemas is still fragmented. In [11], Jurgen Van Ham presents JEScala, a language that captures coordination schemas in a more expressive and modular way by leveraging a seamless integration of an advanced event system with join abstractions. The implementation of joins-based state machines is discussed with alternative faster implementations made possible through a domain specific language. Event monitors are introduced as a way of synchronizing event handling and building concurrent event-based applications from sequential event-based parts.

6.2.2.3. Model Lazy Transformation

The Object Constraint Language (OCL) is a central component in modeling and transformation languages such as the Unified Modeling Language (UML), the Meta Object Facility (MOF), and Query View Transformation (QVT). OCL is standardized as a strict functional language. We have proposed a lazy evaluation strategy for OCL [36]. This lazy evaluation semantics is beneficial in some model-driven engineering scenarios for speeding up the evaluation times for very large models, simplifying expressions on models by using infinite

data structures (e.g., infinite models) and increasing the reusability of OCL libraries. We have implemented the approach on the ATL virtual machine EMFTVM. This is a joint work with the Inria team Atlanmod.

6.2.2.4. Composition Mechanisms for Constraints Generalization

Structural time series (pattern for sequences of values) can be described with numerous automata-based constraints. In [12], we describe a large family of constraints for structural time series by means of function composition. We formalize the patterns using finite transducers. Based on that description, we automatically synthesize automata with accumulators, as well as constraint checkers. The description scheme not only unifies the structure of the existing 30 time-series constraints, but also leads to over 600 new constraints, with more than 100,000 lines of synthesized code. This is a joint work with the Inria team Tasc.

6.3. Software Composition

Participants: Walid Benghrabit, Ronan-Alexandre Cherrueau, Rémi Douence, Hervé Grall, Jean-Claude Royer, Mario Südholt.

6.3.1. Constructive Security

Nowadays we are witnessing the wide-spread use of cloud services. As a result, more and more end-users (individuals and businesses) are using these services for achieving their electronic transactions (shopping, administrative procedures, B2B transactions, etc.). In such scenarios, personal data is generally flowing between several entities and end-users need (i) to be aware of the management, processing, storage and retention of personal data, and (ii) to have necessary means to hold service providers accountable for the usage of their data. Usual preventive security mechanisms are not adequate in a world where personal data can be exchanged on-line between different parties and/or stored at multiple jurisdictions. Accountability becomes a necessary principle for the trustworthiness of open computer systems. It regards the responsibility and liability for the data handling performed by a computer system on behalf of an organization. In case of misconduct (e.g. security breaches, personal data leak, etc.), accountability should imply remediation and redress actions, as in the real life.

In 2015, we have contributed two main results: first, techniques for the logic-based definition, analysis and verification of accountability properties; second, a new framework for the compositional definition of privacy-properties and their type-based enforcement.

6.3.1.1. Logic-based accountability properties

We have proposed a framework for the representation of accountability policies [37]. This framework comes with two novel accountability policy languages; the Abstract Accountability Language (AAL), which is devoted to the representation of preferences/obligations in an human readable fashion, and a concrete one for the mapping to concrete enforceable policies. Our efforts have focused on a formal foundation for the AAL language and some applications.

We have also introduced an approach to assist the design of accountable applications [21]. In particular, we consider an application's abstract component design and we introduce a logical approach allowing various static verification. This approach offers effective means to early check the design and the behavior of an application and its offered/required services. We motivate our work with a realistic use case coming from the A4Cloud project and validate our proposal with experiments using the TSPASS theorem prover. This prover is competitive with other model-checkers and sat solvers and we gain a more abstract approach than with our previous experiment with a model-checker. It makes also easier the link with end users, for instance privacy officers.

To give a formal foundation of the AAL language we define a translation into first-order temporal logic [20]. We introduce a formula to interpret accountability and a natural criterion to achieve the accountability compliance of two clauses. We continue to apply it to an health care system taking into account data privacy features, data transfers and location processing. We demonstrate few heuristics to speed up the resolution time and to assist in conflict detection. Tool support (AccLab) has been provided to support editing, checking and proving AAL clauses.

6.3.1.2. Composition of Privacy-Enforcement Techniques

Today's large-scale computations, e.g., in the Cloud, are subject to a multitude of risks concerning the divulging and ownership of private data. Privacy risks are mainly addressed using a large variety of encryption-based techniques. We have proposed a compositional approach for the declarative and correct composition of privacy-preserving applications in the Cloud [22], [38]. Our approach provides language support for the compositional definition of encryption-based and fragmentation-based privacy-preserving algorithms. This language comes equipped with a set of laws that allows us to verify privacy properties. We have provided implementation support in Scala that ensures certain privacy properties by construction using advanced features of Scala's type system.

6.3.2. Modular systems

6.3.2.1. Modularity for Javascript Interpreters.

With an initial motivation based on the security of web applications written in JavaScript, we have provided new techniques for the instrumentation of an interpreter for a dynamic analysis as a crosscutting concern [31]. We have defined the instrumentation problem — an extension to the expression problem with a focus on modifying interpreters. We have then shown how we can instrument an interpreter for a simple language using only the bare language features provided by JavaScript.

6.4. Cloud applications and infrastructures

Participants: Frederico Alvares, Simon Dupont, Md Sabbir Hasan, Adrien Lebre, Thomas Ledoux, Jonathan Lejeune, Guillaume Le Louët, Jean-Marc Menaud, Jonathan Pastor, Mario Südholt.

In 2015, we have provided solutions for Cloud-based and distributed programming, virtual environments and data centers.

6.4.1. Cloud and distributed programming

6.4.1.1. Cloud elasticity

Cloud Computing has provided important new means for the capacity management of resources. The elasticity and the economy of scale are the intrinsic elements that differentiate it from traditional computing paradigm.

A good capacity planning method is a necessary factor but not sufficient to fully exploit Cloud elasticity. In [26], we propose innovative policies for resource management to achieve the optimal balance between capacity and quality of Cloud services The main idea is to finely control the scalability and the termination of virtual machines with respect to several criteria such as the lifecycle of the instances (*e.g.* initialization time) or their cost. The approach was evaluated on an Amazon EC2 cluster. Experimental results illustrate the soundness of the proposed approach and the impact of scalability/termination resource policies: a cost saving of as much as 30% can be achieved with a minimal number of violations, as small as 1%.

In order to improve Cloud elasticity, we advocate that the software layer can take part in the elasticity process as the overhead of software reconfiguration can be usually considered negligible compared to infrastructural costs. Thanks to this extra level of elasticity, we are able to define cloud reconfigurations that enact elasticity in both the software and infrastructure layers. In [23], we present an autonomic approach to manage cloud elasticity in a cross-layered manner. First, we enhance cloud elasticity with the software elasticity model. Then, we describe how our autonomic cloud elasticity model relies on the dynamic selection of elasticity tactics. We present an experimental analysis of a subset of those elasticity tactics under different scenarios in order to provide insights on strategies that could drive the autonomic selection of the proper tactics to be applied.

6.4.1.2. Service-level agreement for the Cloud

Quality-of-service and SLA guarantees are among the major challenges of cloud-based services. In [18], we first present a new cloud model called SLAaaS — SLA aware Service. SLAaaS considers QoS levels and SLA as first class citizens of cloud-based services. This model is orthogonal to other SaaS, PaaS, and IaaS cloud models, and may apply to any of them. More specifically, we make three contributions: (i) we provide a domain-specific language that allows to define SLA constraints in cloud services; (ii) we present a general control-theoretic approach for managing cloud service SLA; (iii) we apply our approach to MapReduce, locking, and e-commerce services.

6.4.1.3. Distributed multi-resource allocation

Generalized distributed mutual exclusion algorithms allow processes to concurrently access a set of shared resources. However, they must ensure an exclusive access to each resource. In order to avoid deadlocks, many of them are based on the strong assumption of a prior knowledge about conflicts between processes' requests. Some other approaches, which do not require such a knowledge, exploit broadcast mechanisms or a global lock, degrading message complexity and synchronization cost. We propose in [29] [41] a new solution for shared resources allocation which reduces the communication between non-conflicting processes without a prior knowledge of processes conflicts. Performance evaluation results show that our solution improves resource use rate by a factor up to 20 compared to a global lock based algorithm.

6.4.2. Virtualization and data centers

In 2015, we have produced results and tools for the simulation of large-scale distributed algorithms, notably VM scheduling algorithms, have contributed new abstractions for storage systems and have devised new means for the introspection of Cloud infrastructures.

6.4.2.1. SimGrid / VMPlaceS

We have developed VMPlaceS [28], a framework providing programming support for the definition of VM placement algorithms, execution support for their simulation at large scales, as well as new means for their trace-based analysis. VMPlaceS enables, in particular, the investigation of placement algorithms in the context of numerous and diverse real-world scenarios. To illustrate relevance of such a tool, we evaluated three different classes of virtualization environments: centralized, hierarchical and fully distributed placement algorithms. We showed that VMPlaceS facilitates the implementation and evaluation of variants of placement algorithms. The corresponding experiments have provided the first systematic results comparing these algorithms in environments including up to one thousand of nodes and ten thousands of VMs in most cases.

While such a number is already valuable and although we finalized the virtualization abstractions in SimGrid [17], we are in touch with the core developers in order to improve the code of VMPlaceS with the ultimate objective of addressing infrastructures up to 100K physical machines and 1 Millions virtual machines over a period of one day.

The current version of VMPlaceS is available on a public git repository :http://beyondtheclouds.github.io/VMPlaceS/.

6.4.2.2. Storage abstractions within the SimGrid framework

With the recent data deluge, storage is becoming the most important resource to master in modern computing infrastructures. Dimensioning and assessing the performance of storage systems are challenges for which simulation constitutes a sound approach. Unfortunately, only a few existing simulators of large scale distributed computing systems go beyond providing merely a notion of storage capacity. In 2015, we contributed to the SimGrid efforts toward the simulation of such systems [27]. Concretely, we characterized the performance behavior of several types of disks to derive a first model of storage resource. This model has been integrated within the SimGrid framework available under the LGPL license (http://simgrid.gforge.inria.fr).

6.4.2.3. Cloud Introspection

Cloud Computing has become a new technical and economic model for many IT companies. By virtualizing services, it allows for a more flexible management of datacenters capacities. However, its elasticity and its flexibility led to the explosion of virtual environments to manage. It's common for a system administrator to manage several hundreds or thousands virtual machines. Without appropriate tools, this administration task may be impossible to achieve.

We purpose in [32] a decision support tool to detect virtual machines with atypical behavior. Virtual machines whose behavior is different from other VMs running in the data center are tagged as atypicals. Our analysis tool is based on a specific partitioning algorithm which identifies VM behaviors. This tool has been validated in production environments and is used by several companies.

To collect finer metrics (for security, energy management etc.), VM introspection an agent can be installed in a VM to intrusively supervise it or the hypervisor can be used to non-intrusively recover the introspection metrics. In the case of intrusive introspection, the agent installed on the VM operating system will retrieve a set of information related to the operating system operation. However, the installation of an agent in the virtual machine increases the cost of deploying the virtual machine and its resource consumption. The Virtual Machine Introspection (VMI) at the hypervisor level (non intrusively) offer a complete, consistent and untainted view of the VM state. This solution allows an isolation of the VMI mechanism from the guest OS, while allowing monitoring and modifying any state of the VM.

We have also provided a comprehensive summary on VM introspection techniques [25]. Existing VMI techniques are analyzed with respect to their approach to closing the "semantic gap" between the (low level) information provided by the hypervisor and the input to the security analysis.

Finally, we have introduced an extension to LibVmi to detect and monitor a process resource consumption inside a VM from the hypervisor [34]. This extension monitor process cpu and ram ressources without probe. This extension can detect abusive cpu resource usage and atypical ram utilization. This fine monitoring system can be used in many context (security, power consumption, fault tolerance).

6.5. Green IT

Participants: Simon Dupont, Md Sabbir Hasan, Thomas Ledoux, Jonathan Lejeune, Guillaume Le Louët, Jean-Marc Menaud.

In 2015, we have provided new models and solutions for the energy-optimal execution of cloud applications in data centers.

6.5.1. Renewable energy

With the emergence of the Future Internet and the emergence of new IT models such as cloud computing, the usage of data centers (DC) and consequently their power consumption increase dramatically. Besides the ecological impact, the energy consumption is a predominant criteria for DC providers since it determines the daily cost of their infrastructure. As a consequence, power management becomes one of the main challenges for DC infrastructures and more generally for large-scale distributed systems.

6.5.1.1. Renewable energy for data centers

We have presented the EPOC project which focuses on optimizing the energy consumption of mono-site DCs connected to the regular electrical grid and to renewable energy sources [19]. A first challenge in this context consists in developing a (for users) transparent distributed system that enables energy-proportional computations from the system to service-oriented levels. The second challenge addresses the corresponding energy issues through collaborative measurements and energy-optimizing actions inside infrastructure-software stack, more precisely between applications and resource management systems. This approach must manage Service Level Agreement (SLA) constraints by striving for the best trade-off between energy cost (from the regular electric grid), its availability (from renewable energy sources), and service degradation (from application reconfiguration issues to job suspension ones). The third challenge embarks pursues energy efficient optical networks as key enablers of the future internet and cloud-networking service deployment through the convergence of optical infrastructure with the upper network layers.

The second challenge is more precisely describe in [30]. In this paper we present PIKA, a framework aiming at reducing the Brownian energy consumption (ie. from non renewable energy sources), and improving the usage of renewable energy for mono-site data centers. PIKA exploits jobs with slack periods, and executes and suspends them depending on the available renewable energy supply. By consolidating the virtual machines (VMs) on the physical servers, PIKA adjusts the number of powered-on servers in order for the overall energy consumption to match the renewable energy supply. Using simulations driven by real-world workloads and solar power traces, we demonstrate that PIKA consumes 41% less Brownian energy and increases 35.3% renewable energy integration ratio in comparison with the baseline algorithm from the literature.

6.5.1.2. Energy monotoring

We have designed SensorScript, a Business-Oriented Domain-Specific Language for Sensor Networks [24], [35]. In smart grids, or more generally the Internet of Things, many research work has been performed on the whole chain, from communication sensors to big data management, through communication middlewares. Few of this work have addressed the problem of gathered data access. In fact, being able, as a system administrator, to manipulate and gather data collected from a set of sensors in a simple and efficient way represents an essential need.

To address this issue, the solution we considered consists of a multi-context modeling for raw data, in the form of a multi-tree: a directed acyclic graph consisting of multiple intricate trees, each of them describing a hierarchy corresponding to a given use context. The objectives are to provide not only a means to rationalize users needs before writing queries, but also to offer a domain-specific language (DSL) which takes advantage of the multi-tree modeling to simplify the experience of pre-identified users that query data.

6.5.1.3. Green SLA and virtualization of green energy

The demand for energy-efficient services is increasing considerably as people are getting more environmentally-conscious in order to build a sustainable society. The main challenge for Cloud providers is to manage Green SLA (Service Level Agreement) constraints for their customers while satisfying their business objectives, such as maximizing profits by lowering expenditure for so-called green (renewable) energy. Since, Green SLA needs to be proposed based on the presence of green energy, the intermittent nature of renewable sources makes it difficult to be achieved. In response, we propose a scheme for green energy management based on three contributions [15]: i) we introduce the concept of virtualization of green energy to address the uncertainty of green energy availability, ii) we extend the Cloud Service Level Agreement (CSLA) language to support Green SLA by introducing two new threshold parameters and iii) we introduce algorithms for Green SLA which leverage the concept of virtualization of green energy to provide interval-specific Green SLA. We have conducted experiments with real workload profiles from PlanetLab and server power model from SPECpower to demonstrate that Green SLA can be successfully established and satisfied without incurring higher cost.

ATLANMODELS Team

7. New Results

7.1. Reverse Engineering & Evolution

Model Driven Reverse Engineering (MDRE), with its applications on software modernization or tool evolution for example, is a discipline in which model-based principles and techniques are used to treat various kinds of (sometimes very large) existing systems. In the continuity the work started several years ago, AtlanMod has been working actively on this research area this year again. The main contributions are the following:

- In the context of the ARTIST FP7 project, the work has been continued on reusing (and extending accordingly) MoDisco and several of its components to provide the Reverse Engineering support required within the project (and more particularly in the context of the use cases provided by our industrial partners). This has been an important year for the team in this project since it successfully ended in November 2015 after final review at the European Commission. At conceptual-level, the proposed overall approach (as a main result of the ARTIST project) and the main lessons we have learned from its application to concrete industrial scenarios have been published and promoted to a large and high-level audience [11]. The ARTIST project in itself, the various research aspects it addresses and the offered technical solutions have also been presented to the Modeling community [22]. At tooling-level, several (MoDisco-based) model discovery components from Java and SQL have been enhanced while made available as part of a second version of the official ARTIST OS Release [0]. A promising work has also been started on studying deeper the automated discovery of behavioral aspects of software applications, notably by working on a pragmatic mapping between a programming language (Java) and a modeling language (the OMG fUML standard) that focuses on executable aspects.
- To facilitate the understanding of existing software applications via the different models describing them, a significant work has been performed related to providing a generic support for dealing with viewpoints and views expressed on set of possibly heterogeneous and large models. To this intent, and directly capitalizing on the work performed in the TEAP FUI project that ended by the end of 2014, the EMF Views prototype has been significantly refined and enhanced with a ViewPoint Definition Language (the VPDL domain-specific language having a SQL-like syntax) notably [18]. Based on this same model viewpoints/views approach, and more particularly on its underlying (meta)model virtualization support, the general problem of lightweight (meta)model extension has been studied more deeply in the context of our work within the MoNoGe FUI project (national). This has already resulted in a corresponding prototype and a DSL for expressing metamodel extensions [17]. Within the coming year, the plan is to continue further this global work on model viewpoints/views in a software understanding and evolution context.
- Software development projects are notoriously complex and difficult to deal with. Several support tools have been introduced in the past decades to ease the development activities such as issue tracking, code review and Source Control Management (SCM) systems. While such tools efficiently track the evolution of a given aspect of the project (e.g., issues or code), they provide just a partial view of the software project and they often lack of querying mechanisms beyond basic support (e.g., command line, simple gui). This is particularly true for projects that rely on Git, one of the most popular SCM systems. Nowadays many tools are built on top of it, however, they do not complement Git with query functionalities and currently none of them proposes a mechanism that unifies the project information scattered in such different tools. In [28], we propose a conceptual schema for Git and an approach that, given a Git repository, exports its data to a relational database in order to (1) promote data integration with other existing Git-based tools relying on databases

[0] http://www.artist-project.eu/tools-of-toolbox/193

and (2) provide query functionalities expressed through standard SQL syntax. To ensure efficiency, our approach comes with an incremental propagation mechanism that refreshes the database content with the latest modifications.

7.2. MDE Scalability

The increasing number of companies embracing MDE methods and tools have exceeded the limits of the current model-based technologies, presenting scalability issues while facing the growing complexity of their data. Since further research and development is imperative in order to maintain MDE techniques as relevant as they are in less complex contexts, we have focused our research in three axes, (i) scalable persistence solutions, (ii) scalable model transformation engines, and (iii) testing of large scale distributed systems.

In [33], we introduce and evaluate a map-based persistence model for MDE tools. We use this model to build a transparent persistence layer for modeling tools, on top of a map-based database engine. The layer can be plugged into the Eclipse Modeling Framework, lowering execution times and memory consumption levels of other existing approaches. Empirical tests are performed based on a typical industrial scenario, model-driven reverse engineering, where very large software models originate from the analysis of massive code bases. The layer is freely distributed and can be immediately used for enhancing the scalability of any existing Eclipse Modeling tool. We learned that—in terms of performance—typical model-access APIs, with fine-grained methods that only allow for one-step-navigation queries, do not benefit from complex relational or graph-based data structures. Much better results are potentially obtained by optimized low-level data structures, like hash-tables, which guarantee low and constant access times. Additional features that may be of interest in scenarios where performance is not an issue (such as versioning and transactional support provided by CDO) have not been considered. In [32] we extend our persistent mechanism to distributed environments by presenting NeoEMF/HBase, a model-persistence backend for the Eclipse Modeling Framework (EMF) built on top of the Apache HBase data store. Model distribution is hidden from client applications, that are transparently provided with the model elements they navigate. Access to remote model elements is decentralized, avoiding the bottleneck of a single access point. The persistence model is based on key-value stores that allow for efficient on-demand model persistence.

Once we develop a high-performance and distributed persistence mechanism for very-large models, we can exploit it to run high-performance computing over such models. One of the central operations in MDE is rule-based model transformation (MT). It is used to specify manipulation operations over structured data coming in the form of model graphs. However, being based on computationally expensive operations like subgraph isomorphism, MT tools are facing issues on both memory occupancy and execution time while dealing with the increasing model size and complexity. One way to overcome these issues is to exploit the wide availability of distributed clusters in the Cloud for the distributed execution of MT. In [24] and [23], we propose an approach to automatically distribute the execution of model transformations written in a popular MT language, ATL, on top of a well-known distributed programming model, MapReduce. We show how the execution semantics of ATL can be aligned with the MapReduce computation model. We describe the extensions to the ATL transformation engine to enable distribution, and we experimentally demonstrate the scalability of this solution in a reverse-engineering scenario.

Another fundamental operation in MDE is model querying. The Object Constraint Language (OCL) is the standard query language proposed by OMG and is a central component in other modeling and transformation languages such as the Unified Modeling Language (UML), the Meta Object Facility (MOF), and Query View Transformation (QVT). OCL is standardized as a strict functional language. In [34], we propose a lazy evaluation strategy for OCL. We argue that a lazy evaluation semantics is beneficial in some model-driven engineering scenarios for: i) lowering evaluation times on very large models; ii) simplifying expressions on models by using infinite data structures (e.g., infinite models); iii) increasing the reusability of OCL libraries. We implement the approach on the ATL virtual machine EMFTVM.

Finally an important class of operations in MDE is bidirectional (i.e. reversible) computation. Especially bidirectional model transformation is a key technology when two models that can change over time have to be kept constantly consistent with each other. In Hidaka et al. we clarify and visualize the space of design

choices for bidirectional transformations from an MDE point of view, in the form of a feature model. The selected list of existing approaches are characterized by mapping them to the feature model. Then the feature model is used to highlight some unexplored research lines in bidirectional transformations, especially in the scalability of such systems.

7.3. Software Quality

We initiated a new line of research in order to investigate Novelty Search (NS) for the automatic generation of test data, in collaboration with the DiverSE team. Our goal is to explore the huge space of test data within the input domain. In this approach, we select test data based on a novelty score showing how different they are compared to all other solutions evaluated so far [25], [26].

CIDRE Project-Team

7. New Results

7.1. Intrusion detection

7.1.1. Alert Correlation in Distributed Systems

In large systems, multiple (host and network) Intrusion Detection Systems (IDS) and many sensors are usually deployed. They continuously and independently generate notifications (event's observations, warnings and alerts). To cope with this amount of collected data, alert correlation systems have to be designed. An alert correlation system aims at exploiting the known relationships between some elements that appear in the flow of low level notifications to generate high semantic meta-alerts. The main goal is to reduce the number of alerts returned to the security administrator and to allow a higher level analysis of the situation. However, producing correlation rules is a highly difficult operation, as it requires both the knowledge of an attacker, and the knowledge of the functionalities of all IDSes involved in the detection process. In [59], [38], [19], we focus on the transformation process that allows to translate the description of a complex attack scenario into correlation rules and its assessment. We show that, once a human expert has provided an action tree derived from an attack tree, a fully automated transformation process can generate exhaustive correlation rules that would be tedious and error prone to enumerate by hand. The transformation relies on a detailed description of various aspects of the real execution environment (topology of the system, deployed services, etc.). Consequently, the generated correlation rules are tightly linked to the characteristics of the monitored information system. The proposed transformation process has been implemented in a prototype that generates correlation rules expressed in an attack description language called Adele. Additionally, a work has been performed to assess the approach on real environment, and to evaluate the accuracy of the rules built.

In the context of the PhD of Mouna Hkimi, we propose a approach to detect intrusions that affect the behavior of distributed applications. To determine whether an observed behavior is normal or not (occurence of an attack), we rely on a model of normal behavior. This model has been built during an initial training phase. During this preliminary phase, the application is executed several times in a safe environment. The gathered traces (sequences of actions) are used to generate an automaton that characterizes all these acceptable behaviors. To reduce the size of the automaton and to be able to accept more general behaviors that are close to the observed traces, the automaton is transformed. These transformations may lead to introduce unacceptable behaviors. Our curent work aims at identifying the possible errors tolerated by the compacted automaton.

7.1.2. Android Malware Analysis

We explore how information flows induced by a tainted application are helpful to understand how this tainted application interacts within other components inside the operating system. For that purpose, we have defined a new data structure called System Flow Graph representing in a graph how a marked data is disseminated (inside the operating system). We have shown that this data structure is helpful to understand and represent malicious behaviors [31]. Our main challenge is thus to be able to produce relevant graphs which means being able to really observe malicious executions.

For that purpose we developed GroddDroid [25] a tool dedicated to the automatic triggering of Android malware. GroddDroid makes a first static analysis of the application bytecode. During this analysis, GroddDroid identifies the suspicious parts of the bytecode and modifies the bytecode in order to exhibit an execution path that leads to these suspicious parts. The application is later reconstructed/recompiled in order to be executed. This way, GroddDroid offers a way to force the suspicious code to be executed and then observed.

7.1.3. Comparative Study of Alert Formats

In the context of the SECEF project, we conduced a comparative study of different existing alert formats [39]. We analyzed two proprietary formats, CEF (HP ArcSight) and LEEF (IBM QRADAR), as well as 4 standard formats, IDMEF (IETF), CEE (MITRE), CIM and CADF (DMTF). We proposed several metrics to compare them based on an accurate revue of every fields proposed by each format. The results show that IDMEF is the most expressive and structured format. However, some fields proposed by other formats are not covered in IDMEF. We proposed some modification of the alert format to take those limitations into account.

7.1.4. Visualization

This year, research on viusalization for security was oriented towards two objectives. First, as we did during the previous years, we tried to provide solution for security analysts to better analyze *a posteriori* events related to security that are happening on a system. Christopher Humphries, who was the first CIDre Ph.D. student on this topic defended his Ph.D. Thesis *User-Centered Security Events Visualization* this December. We should also mention that we presented a prototype of our tool ELVIS during the FIC 2015 in Lille on the Pôle Cyber-Défense area.

This year, we also started research on a new topic in visualization for security. By contrast with our previous work that was dedicated to forensics, i.e. *a posteriori* analysis of security events, we started this year to study real time analysis of alerts generated by an IDS. The idea here is to allow better monitoring of what is currently happening on a system. We proposed VEGAS, a tool that allows front-line security operators to perform a first triage of the alerts to provide consistent groups of alerts to security analysts. A new Ph.D. student, Damien Crémilleux, was hired on a DGA-MI funding, to work on this topic. VEGAS was presented during the poster session of VizSec 2015 [58] that took place in Chicago, Illinois, USA on the 26th of October 2015.

7.2. Privacy

7.2.1. The Right to be Forgotten

The right to be forgotten, or to oblivion, is an aspect of privacy rights. It relates to the need for individuals to be able to leave a part of their past behind them, to change their mind about something or to take a new start in a given domain. The final report of the DAO project [53] presents an analysis of the concept from a multidisciplinary point of view, including a sociological study, a legal state of the art assorted with insights of possible evolutions, and a technical state of the art along with the proposal of a new architecture [22]. A joint technical and legal analysis of the conceptual and technical issues specific to social networks is also proposed. From the point of view of a computer scientist, the most obvious issue with the right to be forgotten is the ability to control the deletion of a piece of information once it has been disclosed and disseminated. In the general case, no guarantees can be provided, but under certain conditions it is possible to enforce remote deletion with reasonable guarantees. In general, it implies that architectural and applicative choices are made beforehand, either to allow for future decisions regarding data made available in a controlled framework, like late modifications of its access policy or the triggering of its destruction, or to plan deletion from the beginning and set a time-to-leave when disclosing the data within a particular environment, or . The approach designed in CIDRE, relying on both ephemeral publication and data degradation techniques, falls in the latter category, improving the utility for third parties (when compared to existing ephemeral publication techniques) and building a new trade-off with the users' privacy needs, by making different versions of the original data, more or less precise, available for different durations, the more detailed information being lost the quickliest.

CIDRE also contributes, through the B<>com IRT, to the supervision (by Annie Blandin, professor at Télécom Bretagne, and Guillaume Piolle) of Gustav Malis's doctoral work in law in the domain of a restrictive case of the right to be forgotten. In this context, very original contributions have been made at the intersection between the two fields. In particular, a joint analysis has been proposed on the roles of legal and computing tools for the implementation of the right to be forgotten [50]. In particular, it seems that the two domains consider the issue with very different perspectives: the computer scientist almost takes for granted that he cannot rely on regulations and on "security through legality", hence the tools he designs are intended to directly empower the

user, putting him in control of his data by using preventive protection techniques. The tools may fail though, or more likely their applicability conditions may not suit all scenarios. When issues arise they may be captured by the regulatory framework, which intends to provide means for reparation and restoration. Both approaches fail to encompass all possible situations and to solve all potential issues, but they provide users and citizens with complementary tools.

The work combining computer science and law conducted in the DAO projet as well as the main conclusions of the project have also been presented in interdisciplinary colloquium by Sébastien Gambs and Maryline Boizard [48], [47].

7.2.2. Private and Secure Location-based Services

Mobility has always been an important aspect of human activities. Nowadays problems of congestion in urban areas due to the massive usage of cars, last-minutes travel needs and progress in information and communication technologies encourage the rise of new transport modes. Among those are carpooling services, which let car owners share the empty seats of their cars with other travellers having the same travel destination. However, the way carpooling services are implemented today raises several privacy issues. In a recent paper, together with researchers from LAAS-CNRS we have proposed to use privacy enhancing technologies to improve the quality of carpooling services by specially taking in consideration privacy aspects [46].

In addition, publishing directly human mobility data raises serious privacy issues due to its inference potential, such as the (re-)identification of individuals. To address these issues and to foster the development of such applications in a privacy-preserving manner, we propose in a recent paper [26] a novel approach in which Call Detail Records (CDRs) are summarized under the form of a differentially-private Bloom filter for the purpose of privately estimating the number of mobile service users moving from one area (region) to another in a given time frame. Our sanitization method is both time and space efficient, and ensures differential privacy while solving the shortcomings of a solution recently proposed. We also report on experiments conducted using a real life CDRs dataset, which show that our method maintains a high utility while providing strong privacy.

Finally, in authentication protocols, a relay attack allows an adversary to impersonate a legitimate prover, possibly located far away from a verifier, by simply forwarding messages between these two entities. The effectiveness of such attacks has been demonstrated in practice in many environments, such as ISO 14443-compliant smartcards and car-locking mechanisms. Distance-bounding (DB) protocols, which enable the verifier to check his proximity to the prover, are a promising countermeasure against relay attacks. In such protocols, the verifier measures the time elapsed between sending a challenge and receiving the associated response of the prover to estimate their proximity. So far, distance bounding has remained mainly a theoretical concept. Indeed in practice, up to our knowledge only three ISO 14443-compliant implementations of DB protocols exist. The first two are implemented on proprietary smartcards while the last one is available on a highly-customized and dedicated hardware. In a recent paper [35], we demonstrated a proof-of-concept implementation of the Swiss-Knife DB protocol on smartphones running in RFID-emulation mode. To our best knowledge, this is the first time that such an implementation has been performed. Our experimental results are encouraging as they show that relay attacks introducing more than 1.5 ms are directly detectable (in general off-the-shelf relay attacks introduce at least 10 ms of delay). We also leverage on the full power of the ISO-DEP specification to implement the same protocol with 8-bit challenges and responses, thus reaching a better security level per execution without increasing the possibility of relay attacks. The analysis of our results leads to new promising research directions in the area of distance bounding.

7.3. Trust

Reputation mechanisms allow users to mutually evaluate their trust. This is achieved through the computation of a reputation score summarizing their past behaviors. Depending on these scores, users are free to accept or refuse to interact with each other. Existing solutions often rely on costly cryptographic tools that may lead to impractical solutions. We have proposed in [41], [40], [28] usable privacy preserving reputation mechanisms. These mechanisms are distributed and handles non-monotonic ratings. Evaluation made on our mechanism reveals it to be fully usable even with cheap on-board computers. This is a very encouraging result as it shows

that privacy does not impede utility and accuracy. This has been achieved by combining distributed algorithms and cryptographic schemes. Our mechanism is independent of the reputation model, that is, our system can integrate any reputation model, preferably one using both positive and negative ratings.

In a mobile ad hoc network we have also considered the problem of designing a reputation system that allows to update and to propagate the computed reputation scores while tolerating Byzantine failures [42]. Each time a correct node uses directly a service, it can determine by itself the quality of service currently provided. This fresh and valid rating information is broadcast immediately to all its current neighbors. Then, while the mobile node moves, it can receive from other nodes other recommendations also related to the same service. Thus it updates continuously its own opinion. Meanwhile it continues to broadcast this updated information. The freshness and the validity of the received/sent information become questionable. We propose a protocol that allows a node to ignore a second hand information when this information is not fresh or not valid. In particular, fake values provided by Byzantine nodes are eliminated when they are not consistent with those gathered from correct nodes. When the quality of service stabilizes, the correct nodes are supposed to provide quite similar recommendations. In this case, we demonstrate that the proposed protocol ensures convergence to a range of possible reputation scores if a necessary condition is satisfied by the mobile nodes. Simulations are conducted in random mobility scenarios. The results show that our algorithm has a better performance than typical methods proposed in previous works.

7.4. Other Topics Related to Security or Distributed Computing

7.4.1. Detection of distributed denial of service attacks

A Denial of Service (DoS) attack tries to progressively take down an Internet resource by flooding this resource with more requests than it is capable to handle. A Distributed Denial of Service (DDoS) attack is a DoS attack triggered by thousands of machines that have been infected by a malicious software, with as immediate consequence the total shut down of targeted web resources (*e.g.*, e-commerce websites). A solution to detect and to mitigate DDoS attacks it to monitor network traffic at routers and to look for highly frequent signatures that might suggest ongoing attacks. A recent strategy followed by the attackers is to hide their massive flow of requests over a multitude of routes, so that locally, these flows do not appear as frequent, while globally they represent a significant portion of the network traffic. The term “iceberg” has been recently introduced to describe such an attack as only a very small part of the iceberg can be observed from each single router. The approach adopted to defend against such new attacks is to rely on multiple routers that locally monitor their network traffic, and upon detection of potential icebergs, inform a monitoring server that aggregates all the monitored information to accurately detect icebergs [29]. Now to prevent the server from being overloaded by all the monitored information, routers continuously keep track of the c (among n) most recent high flows (modeled as items) prior to sending them to the server, and throw away all the items that appear with a small probability p_i, and such that the sum of these small probabilities is modeled by probability p_0. Parameter c is dimensioned so that the frequency at which all the routers send their c last frequent items is low enough to enable the server to aggregate all of them and to trigger a DDoS alarm when needed. This amounts to compute the time needed to collect c distinct items among n frequent ones. A thorough analysis of the time needed to collect c distinct items appears in [16], [15].

7.4.2. Metrics Estimation on Very Large Data Streams

Huge data flows have become very common in the last decade. This has motivated the design of online algorithms that allow the accurate estimation of statistics on very large data flows. A rich body of algorithms and techniques have been proposed for the past several years to efficiently compute statistics on massive data streams. In particular, estimating the number of times data items recur in data streams in real time enables, for example, the detection of worms and denial of service attacks in intrusion detection services, or the traffic monitoring in cloud computing applications. Two main approaches exist to monitor in real time massive data streams. The first one consists in regularly sampling the input streams so that only a limited amount of data items is locally kept. This allows to exactly compute functions on these samples. However, accuracy of this computation with respect to the stream in its entirety fully depends on the volume of data items that has been

sampled and their order in the stream. In contrast, the streaming approach consists in scanning each piece of data of the input stream on the fly, and in locally keeping only compact synopses or *sketches* that contain the most important information about these data. This approach enables us to derive some data streams statistics with guaranteed error bounds without making any assumptions on the order in which data items are received at nodes. Sketches highly rely on the properties of hashing functions to extract statistics from them. Sketches vary according to the number of hash functions they use, and the type of operations they use to extract statistics. The *Count-Min sketch* algorithm proposed by Cormode and Muthukrishnan in 2005 so far predominates all the other ones in terms of space and time needed to guarantee an additive ε-accuracy on the estimation of item frequencies. Briefly, this technique performs t random projections of the set of items of the input stream into a much smaller co-domain of size k, with $k = \lceil e/\epsilon \rceil$ and $t = \lceil \log(1/\delta) \rceil$ in which $0 < \epsilon, \delta < 1$. The user defined parameters ϵ and δ represent respectively the accuracy of the approximation, and the probability with which the accuracy holds. However, because k is typically much smaller than the total number of distinct items in the input stream, hash collisions do occur. This affects the estimation of item frequency when the size of the stream is large. In this work, we have proposed an alternative approach to reduce the impact of collisions on the estimation of item frequency. The intuition of our idea is that by keeping track of the most frequent items of the stream, and by removing their weight from the one of the items with which these frequent items collide, the over-estimation of non frequent items is drastically decreased [21].

We have also proposed a metric, called codeviation, that allows to evaluate the correlation between distributed streams [27]. This metric is inspired from classical metric in statistics and probability theory, and as such allows us to understand how observed quantities change together, and in which proportion. We then propose to estimate the codeviation in the data stream model. In this model, functions are estimated on a huge sequence of data items, in an online fashion, and with a very small amount of memory with respect to both the size of the input stream and the values domain from which data items are drawn. We give upper and lower bounds on the quality of the codeviation, and provide both local and distributed algorithms that additively approximates the codeviation among n data streams by using a sublinear number of bits of space in the size of the domain value from which data items are drawn, and the maximal stream length. To the best of our knowledge, such a metric has never been proposed so far.

7.4.3. Stream Processing Systems

Stream processing systems are today gaining momentum as a tool to perform analytics on continuous data streams. Their ability to produce analysis results with sub-second latencies, coupled with their scalability, makes them the preferred choice for many big data companies.

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges. Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operator's ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource utilization and thus larger throughputs and reduced tuple processing latencies. We have proposed a new key grouping technique targeted toward applications working on input streams characterized by a skewed value distribution [44]. Our solution is based on the observation that when the values used to perform the grouping have skewed frequencies, e.g. they can be approximated with a Zipfian distribution, the few most frequent values (the *heavy hitters*) drive the load distribution, while the remaining largest fraction of the values (the *sparse items*) appear so rarely in the stream that the relative impact of each of them on the global load balance is negligible. We have shown, through a theoretical analysis, that our solution provides on average near-optimal mappings using sub-linear space in the number of tuples read from the input stream in the learning phase and the support (value domain) of the tuples. In particular this analysis presents new results regarding the expected error made on the estimation of the frequency of heavy hitters.

7.4.4. Randomized Message-Passing Test-and-Set

In [30], we have presented a solution to the well-known Test&Set operation in an asynchronous system prone to process crashes. Test&Set is a synchronization operation that, when invoked by a set of processes,

returns yes to a unique process and returns no to all the others. Recently many advances in implementing Test&Set objects have been achieved, however all of them target the shared memory model. In this paper we propose an implementation of a Test&Set object in the message passing model. This implementation can be invoked by any number p $\leq$ n of processes where n is the total number of processes in the system. It has an expected individual step complexity in $O(\log p)$ against an oblivious adversary, and an expected individual message complexity in $O(n)$. The proposed Test&Set object is built atop a new basic building block, called selector, that allows to select a winning group among two groups of processes. We propose a message-passing implementation of the selector whose step complexity is constant. We are not aware of any other implementation of the Test&Set operation in the message passing model.

7.4.5. Population Protocol Model

The population protocol model, introduced by Angluin et his colleagues in 2006, provides theoretical foundations for analyzing global properties emerging from pairwise interactions among a large number of anonymous agents. In the population protocol model, agents are modeled as identical and deterministic finite state machines, *i.e.* each agent can be in a finite number of states while waiting to execute a transition. When two agents interact, they communicate their local state, and can move from one state to another according to a joint transition function. The patterns of interaction are unpredictable, however they must be fair, in the sense that any interaction that should possibly appear cannot be avoided forever. The ultimate goal of population protocols is for all the agents to converge to a correct value independently of the interaction pattern. Examples of systems whose behavior can be modeled by population protocols range from molecule interactions of a chemical process to sensor networks in which agents, which are small devices embedded on animals, interact each time two animals are in the same radio range.

In this work, we focus on an quite important related question. Namely, is there a population protocol that exactly counts the difference κ between the number of agents that initially set their state to A and the one that initially set it to B, and can it be solved in an efficient way, that is with the guarantee that each agent should converge to the exact value of κ after having triggered a sub-linear number of interactions in the size of the system [43].

We answer this question by the affirmative by presenting a $O(n^{3/2})$-state population protocol that allows each agent to converge to the exact solution by interacting no more than $O(\log n)$ times. The proposed protocol is very simple (as is true for most known population protocols), but is general enough to be used to solve different types of tasks.

DIONYSOS Project-Team

6. New Results

6.1. Quality of Experience

Participants: Yassine Hadjadj-Aoul, Gerardo Rubino.

QoE in mobile networks. We consider in [43] an important Quality of Experience (QoE) indicator in cellular networks that is reneging of users due to impatience. We specifically consider a cell under heavy load conditions, modeled as a multiclass Processor Sharing system, and compute the reneging probability by using a fluid limit analysis. In order to enhance the user QoE, we propose a radio resource allocation control scheme that minimizes the global reneging rates. This control scheme is based on the α-fair scheduling framework and adapts the scheduler parameter depending on the traffic load. While the proposed scheme is simple, our results show that it achieves important performance gains. This work is extended in [42]. By solving the fixed point equation, we obtain a new QoE perturbation metric quantifying the impact of reneging on the performance of the system. This metric is then used to devise a new pricing scheme accounting of reneging. We specifically propose several flavors of this scheme around the idea of having a flat rate for accessing the network and an elastic price related to the level of QoE perturbation induced by communications.

In order to offer a high media quality and a good user satisfaction, the media streaming service requires that transport protocols can be adapted continuously to the network parameters. However, the diversity of terminals (e.g., tablets, smart phones, laptops) and their corresponding capabilities, mean that users' agnostic solutions are inefficient to cope with such diverse contexts. Indeed, the intrinsic characteristics and parameters of the terminals (i.e., devices) need to be taken into account on the video streaming adaptation process. In [17], we propose an adaptive video streaming solution to improve the user satisfaction factor by adapting the TCP parameters according to the user's parameters on mobile networks. The user satisfaction factor is calculated according to some metrics driven from the user's quality of experience (QoE). The work is validated through our proposal based on a new mobile agent developed on a Linux script platform and tested on different kinds of devices with different scenarios.

Learning tools. Our QoE measuring techniques (see 3.2) are based on statistical learning methods, and we have been using Random Neural Networks as our main learning tool. These are actually open queueing networks where customers have a "sign" and behave analogously as neural spiking signals. They have been proposed by Gelenbe in the 80s, and have been used in many areas since then. In [26], we published a survey about the tool, where we develop in some detail their use in supervised learning, not only for the case of interest in PSQA, our QoE measuring technology. We also discuss the use of powerful optimization methodology, first and second order techniques, that have proved to be very effective in the standard Neural Network area.

Recently, we started to explore new learning techniques. The first reason is not the search for more accurate tools, because ours are, we claim, as accurate as they can be, it is to improve robustness. The second reason is to extend our QoE measuring tools to richer contexts, mainly when we take into account time, that is, time series data. This comes from the observation that in many cases, the way people perceive quality has some "inertia" and depends on the quality perceived some minutes ago. In [66] we explored the capabilities of a recently proposed method called "Reservoir Computing (RC) with Random Static Projections" which combines two ideas, the now classic Reservoir Computing approach and Extreme Learning Machines (ELMs). In our paper, we replaced the ELMs by Radial Basis Functions (RBF) projections. We illustrated the good behavior of this variation of the original technique basically using known benchmarks.

In [67], we perform a detailed analysis of one of the main instances of the Reservoir Computing idea, called Echo State Network (ESN). This type of model has several parameters to adjust, that have an impact on the performances of the learning procedure. For instance, it has been shown that the spectral radius of the reservoir matrix (the recurrent network structure that doesn't learn during the process) is related to the accuracy and the memory capabilities of the technology. The size of the reservoir is also a parameter to adjust when configuring

an ESN for performing some specific task. One of the results of our work is the fact that the periodic or pseudo-period nature of data is also an important factor to be taken into account when designing an ESN, since it has an influence on the impact of parameters such as the previously mentioned spectral radius.

QoE and emergency management. As a by-product of our activities around QoE, we started to work on an application where, instead of evaluating the QoE of, say, a video or voice application, we wanted to evaluate the way users perceive a service not necessarily based on audio or video content. This was related to our participation to the European project QuEEN (see 8.2.2). We finished by building a platform where we test different ideas for managing an emergency situation. In our system, we include an automatic evaluator of the perceived quality of the related voice and video communications, since in the case of some catastrophes, the communications can be seriously damaged and it is critical to automatically detect the issue in order to report the problem and to take appropriate countermeasures, when possible. In [55], we describe some of the aspects of our system and of the implemented mechanisms, and we present some design problems and their solutions, together with illustrations of the capabilities of the tool.

6.2. Analytic models

Participants: Gerardo Rubino, Bruno Sericola.

Sojourn times in Markovian models. In [74], we discuss different issues related to the time a Markov chain spends in a part of its state space. This is relevant in many application areas including those interesting Dionysos, namely, performance and dependability analysis of complex systems. For instance, in dependability, the reliability of a system subject to failures and repairs of its components, is, in terms of a discrete-space model of it, the probability that it remains in the subset of operational or up states during the whole time interval $[0, t]$. In performance, the occupancy factor of some server is the probability that, in steady state, the model belongs to the subset of states where the server is busy. This book chapter reviews some past work done by the authors on this topic, and add some new insights on the properties of these sojourn times.

Queuing systems in equilibrium. In the late 70s, Leonard Kleinrock proposed a metric able to capture the tradeoff between the work done by a system and its cost, or, in terms of queueing systems, between throughput and mean response time. The new metric was called *power* and among its properties, it satisfies a nice one informally called "keep the pipe full", specifying that the operation point of some queues (mainly the $M/M/1$ one) giving the maximal possible value to the power is when the mean backlog is 1. In [56], we took back this idea to explore what happens when we consider Jackson queuing networks. After showing that the same property holds for them and exploring other ones, we show that the power metric has some drawbacks when considering multiserver queues and networks of queues. We then propose a new metric that we called *effectiveness*, identical to power when there is a single queue with a single server, but different otherwise, that avoids these drawbacks. We analyze it and, in particular, we show that the same "keep the pipe full" holds for it.

Transient analysis of queuing systems. In a well-known book [86], today out of press, a concept of dual of a birth-and-death process is proposed, based on stochastic monotonicity. In past work [88] we showed that this concept coupled with the classical randomization or uniformization of continuous time Markov chains and lattice path combinatorics, allowed to derive analytical expressions of the transient distribution of several Markovian queueing systems. Recently, we discovered two new things: first, that this dual concept can be generalized to arbitrary systems of ordinary differential equations (ODEs) and still keep its main properties; second, that we can define a similar transformation than uniformization, that can be applied to arbitrary systems of ODEs and again, holding similar properties than the former. We respectively called pseudo-dual and pseudo-randomization the two concepts and associated methods. In [69], we presented these ideas and first results about them. We illustrated their use, and how they allow to obtain analytical expressions of transient queues' distributions in cases where Anderson's dual doesn't exist (see [87].

In [68], we present results concerning some aspects of the behavior of a queueing system observed during a fixed time period of the form $[0, t]$. The two aspects we looked at in this work are the loss process of a finite capacity model during the considered $[0, t]$, and the maximal backlog reached at a queue over the interval.

Following the classical procedure mentioned below, consisting in using uniformization to go to discrete time and then, combinatorial techniques, we develop numerical schemes to analyze both aspects of some basic queueing systems.

Network reliability. In [28], we consider the classical network design "Capacitated m-Ring Star Problem" (CmRSP), where we look for m rings connecting two nodes in a network at minimum cost. We add to this model the fact that links can fail, and propose a new paradigm that we call "Capacitated m-Ring Star Problem with Diameter Constrained Reliability " (in short, CmRSP-DCR), where we look again for a minimal cost spanning graph of the set of nodes in the network that connects the selected source and terminal, *while satisfying a Diameter Constrained Reliability (DCR) condition*. The DCR is the probability that the two nodes can communicate by means of paths having lengths bounded by some fixed value d. We prove that this problem is NP-hard, and we propose a GRASP-based approach to solve it.

Fluid models. In [19] we study congestion periods in a finite fluid buffer when the net input rate depends upon a recurrent Markov process; congestion occurs when the buffer content is equal to the buffer capacity. We consider the duration of congestion periods as well as the associated volume of lost information. We derive their distributions in a typical stationary busy period of the buffer. Our goal is to compute the exact expression of the loss probability in the system, which is usually approximated by the probability that the occupancy of the infinite buffer is greater than the buffer capacity under consideration. Moreover, by using general results of the theory of Markovian arrival processes, we show that the duration of congestion and the volume of lost information have phase-type distributions.

6.3. Performance Evaluation of Distributed Systems

Participants: Bruno Sericola, Yann Busnel, Pierre L'Ecuyer.

Detection of distributed deny of service attacks. A Deny of Service (DoS) attack tries to progressively take down an Internet resource by flooding this resource with more requests than it is capable to handle. A Distributed Deny of Service (DDoS) attack is a DoS attack triggered by thousands of machines that have been infected by a malicious software, with as immediate consequence the total shut down of targeted web resources (*e.g.*, e-commerce websites). A solution to detect and to mitigate DDoS attacks it to monitor network traffic at routers and to look for highly frequent signatures that might suggest ongoing attacks. A recent strategy followed by the attackers is to hide their massive flow of requests over a multitude of routes, so that locally, these flows do not appear as frequent, while globally they represent a significant portion of the network traffic. The term "iceberg" has been recently introduced to describe such an attack as only a very small part of the iceberg can be observed from each single router. The approach adopted to defend against such new attacks is to rely on multiple routers that locally monitor their network traffic, and upon detection of potential icebergs, inform a monitoring server that aggregates all the monitored information to accurately detect icebergs [36]. Now to prevent the server from being overloaded by all the monitored information, routers continuously keep track of the c (among n) most recent high flows (modeled as items) prior to sending them to the server, and throw away all the items that appear with a small probability. Parameter c is dimensioned so that the frequency at which all the routers send their c last frequent items is low enough to enable the server to aggregate all of them and to trigger a DDoS alarm when needed. This amounts to compute the time needed to collect c distinct items among n frequent ones. A thorough analysis of the time needed to collect c distinct items appears in [12], [11].

Stream Processing Systems. Stream processing systems are today gaining momentum as tools to perform analytics on continuous data streams. Their ability to produce analysis results with sub-second latencies, coupled with their scalability, makes them the preferred choice for many big data companies.

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges (the arcs). Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operators' ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource

utilization and thus larger throughputs and reduced tuple processing latencies. We have proposed a new key grouping technique targeted toward applications working on input streams characterized by a skewed value distribution [53]. Our solution is based on the observation that when the values used to perform the grouping have skewed frequencies, the few most frequent values (the *heavy hitters*) drive the load distribution, while the remaining largest fraction of the values (the *sparse items*) appear so rarely in the stream that the relative impact of each of them on the global load balance is negligible. We have shown, through a theoretical analysis, that our solution provides on average near-optimal mappings using sub-linear spaces in the number of tuples read from the input stream in the learning phase and the support (value domain) of the tuples. In particular this analysis presents new results regarding the expected error made on the estimation of the frequency of heavy hitters.

Randomized Message-Passing Test-and-Set. In [37], we have presented a solution to the well-known Test&Set operation in an asynchronous system prone to process crashes. Test&Set is a synchronization operation that, when invoked by a set of processes, returns yes to a unique process and returns no to all the others. Recently, many advances in implementing Test&Set objects have been achieved. However, all of them target the shared memory model. In this paper we propose an implementation of a Test&Set object in the message passing model. This implementation can be invoked by any number $p \leq n$ of processes where n is the total number of processes in the system. It has an expected individual step complexity in $O(\log p)$ against an oblivious adversary, and an expected individual message complexity in $O(n)$. The proposed Test&Set object is built atop a new basic building block, called selector, that allows to select a winning group among two groups of processes. We propose a message-passing implementation of the selector whose step complexity is constant. We are not aware of any other implementation of the Test&Set operation in the message passing model.

Population Protocol Model. The population protocol model, introduced by Angluin and his colleagues in 2006, provides theoretical foundations for analyzing global properties emerging from pairwise interactions among a large number of anonymous agents. In the population protocol model, agents are modeled as identical and finite state machines, i.e each agent can be in a finite number of states while waiting to execute a transition. When two agents interact, they communicate their local state, and can move from one state to another according to a transition function. The ultimate goal of population protocols is for all the agents to converge to the same value. Examples of systems whose behavior can be modeled by population protocols range from molecule interactions of a chemical process to sensor networks in which agents, which are small devices embedded for instance in animals, interact each time two animals are in the same radio range.

In this work, we focus on a quite important related question. Namely, is there a population protocol that exactly counts the difference κ between the number of agents that initially set their state to A and the one that initially set it to B, and can it be solved in an efficient way, that is with the guarantee that each agent should converge to the exact value of κ after having triggered a sub-linear number of interactions in the size of the system [49]?. We answer this question by the affirmative by presenting a $O(n^{3/2})$-state population protocol that allows each agent to converge to the exact solution by interacting no more than $O(\log n)$ times. The proposed protocol is very simple (as is true for most known population protocols), but is general enough to be used to solve different types of tasks.

Call centers. We develop research activities around the analysis and design of call centers, from a performance perspective. The effective management of call centers is a challenging task mainly because managers are consistently facing considerable uncertainty. Among important sources of uncertainty are call arrival rates which are typically time-varying, stochastic, dependent across time periods and across call types, and often affected by external events. Accurately modeling and forecasting future call arrival volumes is a complicated issue which is critical for making important operational decisions, such as staffing and scheduling, in the call center. In [20] we review the existing literature on modeling and forecasting call arrivals. We also develop in [58] customer delay predictors for multi-skill call centers that take as inputs the queueing state upon arrival and the waiting time of the last customer served. Barely any predictor currently exists for the multi-skill case. We introduce two new predictors that use cubic regression splines and artificial neural networks, respectively, and whose parameters are optimized (or learned) from observation data obtained by simulation.

6.4. Wireless Networks

Participants: Osama Arouk, Btissam Er-Rahmadi, Adlen Ksentini, Meriem Bouzouita, Pantelis Frangoudis, Yassine Hadjadj-Aoul, Gerardo Rubino.

We are continuing our activities around wireless and mobile networks, by focusing more on leveraging the current mobile and wireless architecture toward building the 5G systems.

LTE improvements. One of the 5G objectives is to support a high number of devices. This not only concerns User Equipment (UE) devices, but also other devices such as sensors and actuators (known also as Internet of Things (IoT)). Sensor and actuator devices communicate generally with a remote server in an automatic way, without any human intervention. This type of communication is known as Machine to Machine (M2M) communication, or Machine Type Communication (MTC). The corresponding traffic is known by its intensity and impact on increasing congestion in both main parts of 4G networks, the Radio Access Network (RAN) and the Core Network. To improve the current LTE system to support MTC, we did several contributions. We proposed in [51] an important enhancement to the Group Paging (GP) mechanism, which is responsible for relaying requests to sensors, in order to gather data. After modeling analytically the GP procedure, we proposed a mechanism that, instead of paging all MTC devices in the same period, calculates the appropriate number of MTCs that reduces the collision probability as well as increases the success probability. In [52], we modeled the Radio Access Channel (RACH) procedure when the MTC devices are activated in a highly synchronized manner during a certain period (synchronized traffic), which is represented by a Beta distribution. The proposed model estimates for each period the exact number of MTC devices that may win the contention.

To control the Random Access Network (RAN) overload and alleviate the access network congestion, 3GPP developed the Access Class Barring (ACB) procedure that depends on an access probability called the ACB factor, without proposing a procedure for calculating such probability. In [72], we have proposed a fluid-based random access model for M2M communications, which was used to determine dynamically the value of the ACB factor that avoids system overload and the radio resources' underutilization at the same time. We proposed in [60] a novel implementation of the ACB mechanism in the context of multiple M2M traffic classes. Based on a scheduling algorithm, we have applied a PID controller to adjust dynamically multiple ACB factors related to each class category, guaranteeing a number of devices around an optimal value that maximizes the Random Access (RA) success probability. In [61], we first present a simple fluid model of MTC devices' random access. This model is then used to derive a novel adaptive regulator of the ACB factor, somehow in contrast with previous existing contributions which generally rely on heuristics. The main advantages of the proposed approach are twofold. First, the proposal is fully compliant with the standard while it reduces significantly the computation and the signaling overheads. Second, it provides an efficient mean to regulate adaptively the ACB factor as it guarantees having an optimal number of MTC devices accessing concurrently to the RAN. The obtained results based on simulations show clearly the robustness of the proposed approach, and its superiority compared to existing proposals.

Another important objective of 5G mobile networks is to accommodate a diverse and ever-increasing number of user equipments (UEs). Coping with the massive signaling overhead expected from UEs is an important hurdle to tackle so as to achieve this objective. In [38], we devised an efficient tracking area list management framework that aims for finding optimal distributions of tracking areas (TAs) in the form of TA lists (TALs) and assigning them to UEs. The objective is to minimize two conflicting metrics: paging overhead and tracking area update (TAU) overhead. We used bargaining games to find the Pareto optimal solution that satisfies both objectives.

WiFi networks improvements. It is well established that WiFi is complementing LTE connections to ensure, wirelessly, high data rate. One idea to improve WiFi towards high data rates is to multiple users' transmissions on both directions, i.e. on the Down Link (DL) and the Up Link (UL). In [50] we devised a novel solution to enhance the TXOP Sharing mechanism, introduced in the 802.11ac amendment, to achieve efficient Down-Link Multi-User Multiple-Input Multiple-Output (DL-MU-MIMO) transmission. First, we give new definitions about both events of successful and failed DL-MU-MIMO transmission. Then, we devise a revised

Backoff procedure for the primary Access Category (AC). In [40] we proposed a novel 802.11ax MAC protocol aiming at reducing the elapsed time in managing the establishment of an UL-MU communication, thus enhancing considerably the system's performance.

On the other hand, the volume of mobile multimedia traffic is fast-growing, challenging the radio and backhaul network infrastructure and calling for alternative content dissemination schemes. To improve user experience and reduce infrastructure load, we exploit implicit social relationships among users and take into account content popularity, proposing push-based prefetching mechanisms which take advantage of the caching and mobile ad hoc networking capabilities of user devices. We use, in [65], bloom filters as summaries of user caches, and design mechanisms to estimate the social distance between users and the popularity of content items, which drive our algorithms. Our simulation-based evaluation shows that our scheme brings caching performance improvements in an order of 10% in terms of absolute cache hit ratio in most of the cases studied, and from 3% to 82% in terms of normalized cache hit ratio gain.

Network selection. With the explosion of mobile data traffic, the Fixed and Mobile Converged (FMC) network are being heavily required. Mobile devices have the capability of connecting to different access networks in the FMC architecture simultaneously. Access network selection becomes an issue when mobile devices are under coverage of different access networks, since a bad selection may lead to network congestion and degrade the QoE of users. In order to address this problem, we model and analyze, in [62] and [63], the interface selection procedure using control theory in the FMC architecture. Based on our model, we designed a controller which can send to mobile devices a network selection command calculated instantly for the access network selection. In [29], we investigated network decentralization in conjunction with the selective IP traffic offload approaches to handle the 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.

Energy efficiency. Due to the ever-growing gap between battery lifetime and hardware/software complexity in addition to application's computing power needs, the energy saving issue becomes crucial. In this context, we proposed, in [13], an end-to-end study of video decoding on different architectures. The study was achieved thanks to a two steps methodology: (1) a comprehensive characterization and evaluation of the performance and the energy consumption of video decoding, (2) an accurate high level energy model based on the characterization step. In [24], we proposed to apply data fragmentation, in slotted CSMA/CA, in a way to allow improving the bandwidth occupation while reducing the latency. We proposed to introduce a network allocation vector (NAV) in the fragmentation mechanism to reduce energy consumption in IEEE 802.15.4. A Markov chain-based analytical model of the fragmentation mechanism was given as well as an analytical model of the energy consumption using a NAV. The analytical results show that the fragmentation technique improves at the same time the throughput, the access delay and the bandwidth occupation. They also show that the NAV mechanism reduces energy consumption when applying the fragmentation technique in slotted CSMA-CA for IEEE 802.15.4.

6.5. Future networks and architectures

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

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 in [70] an automated hierarchical controller-based architecture handling the whole control chain.

Mobile cloud. One of the 5G-architecture visions considers the usage of cloud to ease mobile networks evolution towards more flexibility and elasticity for handling resources; building the concept of carrier cloud. Software Defined Networking (SDN) and Network Function Virtualization (NFV) represent the key enabler of carrier cloud. In [57], we addressed the problem of Virtual Network Function (VNF) placement in the carrier

cloud. Indeed, we proposed a placement solution that has two main design goals: i) minimizing path between users and their respective data anchor gateways and ii) optimizing their sessions? mobility. The two design goals effectively represent two conflicting objectives that we deal with considering the mobility features and service usage behavioral patterns of mobile users, in addition to the mobile operators? cost in terms of the total number of instantiated VNFs to build a Virtual Network Infrastructure (VNI). We modeled this problem using an optimization formulation having these conflicting objectives, and then used Bargaining Game to find the Pareto optimal solution. We are continuing our improvement to the Follow Me Cloud (FMC), which was devised by our team conjointly with NEC labs. In [33], we proposed a FMC architecture that relies on PMIPv6 to handle mobility, and SDN to update the flow table of the anchor routers when a service has moved from one Data Center to another. In [10] and [32], we addressed the challenge of flow table scalability problem, which may arise in FMC to high number of mobile users. To this aim, we proposed a two-level hierarchical SDN controllers architecture in order to distribute the SDN/OpenFlow control plane. Another objective of 5G is to reduce network latency to 1ms, which will ease computation offloading. Thus, it will be possible to run applications on UE device, even if the latter has low computation capability, by offloading part of the code to a remote server. In [44], we were interested on studying the opportunities to offload part of one of the well known game engine in the literature, i.e. Unity 3D. We built a data set representing the CPU-GPU use of several games; allowing us to understand which modules might be offloaded to a remote server in the Mobile Cloud.

6.6. Network Economics

Participant: Bruno Tuffin.

The general field of network economics, analyzing the relationships between all acts of the digital economy, has been an important subject for years in the team. The whole problem of network economics, from theory to practice, describing all issues and challenges, is described in our book “Telecommunication Network Economics From Theory to Applications” (P. Maillé and B. Tuffin, Cambridge U. Press, 2014.

Network neutrality. Among the topics we have particularly focused on, the network neutrality debate was a major concern in 2015. In [23], [80], [83] we recall the debate and highlight the fact that neutrality principles can be bypassed in many ways without violating the rules currently evoked in the debate. For example via Content Delivery Networks (CDNs), which deliver content on behalf of content providers for a fee, or via search engines, which can hinder competition and innovation by affecting the visibility and accessibility of content. In [23], we challenge the definition of net neutrality as it is generally discussed. Our goal there is to initiate a relevant debate for net neutrality in an increasingly complex Internet ecosystem, and to provide examples of possible neutrality rules for different levels of the delivery chain, this level separation being inspired by the OSI layer model.

As particular ways to bypass the current neutrality principles, we have particularly focused on CDNs. We for example investigate in [47] the impact of decisions made by a CDN willing to maximize its revenue through the management of cache servers. Based on a model with two network providers, we highlight that revenue-oriented management policies can affect the user-perceived quality of experience, impacting the competition among network access providers in favor of the largest one. Since this contradicts the principle underpinning network neutrality?although not with the technical net neutrality rules?we discuss the necessity to regulate CDN activity. Also, one of the main argument toward neutrality being that it favors innovation, we study in [46] the impact of CDNs" activity on other actors of the supply chain. Our findings indicate that vertically integrating a CDN helps Internet Service Providers (ISPs) collect fees from Content Providers (CPs), hence circumventing the interdiction of side payments coming from net neutrality rules. However, this outcome is socially much better in terms of user quality and innovation fostering than having separate actors providing the access and CDN services: in the latter case double marginalization (both ISP and CDN trying to get some value from the supply chain) leads to suboptimal investments in CDN storage capacities and higher prices for CPs, resulting in reduced innovation.

Another model we have developed is for understanding the behavior of some big providers actually paying side payment to ISPs while still officially in favor of neutrality. To better understand this strategical behavior, we have presented a simple model in [59] providing some insight on whether or not paying side payments for an incumbent provider is a way to create barriers to entry for competitors. It also investigates the economic consequences on all actors: incumbent and new entrant content providers, users, and the Internet Service Provider.It then describes how the side payment can be determined as a Nash bargaining solution.

Pricing access networks. Access networks in a competitive context has been a topic of research for a while. In the Internet, the data charging scheme has usually been flat rate. But more recently, especially for mobile data traffic, we have seen more diversity in the pricing offers, such as volume-based ones or cap-based ones. We study in [48] the behavior of heterogeneous users facing two offers: a volume-based one and a flat-rate one. On top of that selection, we investigate 1) the relevance for an ISP to propose the two types of offers, and optimize the corresponding prices, and 2) the existence of a solution to the pricing game when the offers come from competing providers.

Sponsored auctions. Advertisement in dedicated webpage spaces or in search engines sponsored slots is usually sold using auctions, with a payment rule that is either per view or per click. But advertisers can be both sensitive to being viewed (brand awareness effect) and being clicked (conversion into sales). In [84], we generalize the auction mechanism by including both pricing components: the displayed advertisers are charged when their ad is displayed, and pay an additional price if the ad is clicked. Applying the results for Vickrey-Clarke-Groves (VCG) auctions, we show how to compute payments to ensure incentive compatibility from advertisers as well as maximize the total value of the advertisement slot(s). We provide tight upper bounds for the loss of efficiency due to applying only pay- per-click (or pay-per-view) pricing instead of our scheme. Those bounds depend on the joint distribution of advertisement visibility and population likelihood to click on ads, and can help identify situations where our mechanism yields significant improvements. We also describe how the commonly used generalized second price (GSP) auction can be extended to this context.

6.7. Monte Carlo

Participants: Pierre L'Ecuyer, Gerardo Rubino, Bruno Tuffin.

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, to accelerate the occurrence of the event and get unbiased estimators of the event of interest with a sufficiently small relative variance. This is the main problem in the area. Dionysos' work focuses then on dealing with the rare event situation. For example, [39] presents an exponential tilting method for exact simulation from the truncated multivariate student-t distribution in high dimensions as an alternative to approximate Markov Chain Monte Carlo sampling.

A non-negligible part of our activity on the application of rare event simulation was about the evaluation of static network reliability models. Our paper [16] focuses on a technique known as Recursive Variance Reduction (RVR) which approaches the unreliability by recursively reducing the graph from the random choice of the first working link on selected cuts. This previously known method is shown to not verify the bounded relative error (BRE) property as reliability of individual links goes to one, i.e., 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 IS technique. Two new estimators are presented in the paper: 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, combines RVR and our zero-variance IS approximation. We show that in both cases BRE property is verified and, moreover, that a vanishing relative error (VRE) 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. In [54], we explore the use of the same powerful RVR idea, but applied in a very general

context, where the system is model by a monotone structure function. In the paper, we illustrate the approach with a very widely used model, a series of k-out-of-m modules.

In a static network reliability model one typically assumes that the failures of the components of the network are independent. This simplifying assumption makes it possible to estimate the network reliability efficiently via specialized Monte Carlo algorithms. Hence, a natural question to consider is whether this independence assumption can be relaxed, while still attaining an elegant and tractable model that permits an efficient Monte Carlo algorithm for unreliability estimation. In [14] we provide one possible answer by considering a static network reliability model with dependent link failures, based on a Marshall-Olkin copula, which models the dependence via shocks that take down subsets of components at exponential times, and propose a collection of adapted versions of permutation Monte Carlo (PMC, a conditional Monte Carlo method), its refinement called the turnip method, and generalized splitting (GS) methods, to estimate very small unreliabilities accurately under this model. The PMC and turnip estimators have bounded relative error when the network topology is fixed while the link failure probabilities converge to 0, whereas GS does not have this property. But when the size of the network (or the number of shocks) increases, PMC and turnip eventually fail, whereas GS works nicely (empirically) for very large networks, with over 5000 shocks in our examples. In [41] we focus on a method proposed by Fishman making use of bounds on the structure function describing in terms of configurations of (independent) link states if the considered nodes are connected. The bounds are based on the computation of (independent) mincuts disconnecting the set of nodes and (independent) minpaths ensuring that they are connected. We analyze here the robustness of the method when the unreliability of links goes to zero. We show that the conditions provided by Fishman are based on a bound and are therefore only sufficient, and provide more insight and examples on the behavior of the method.

PMC is an effective way of estimating the unreliability of a static network when this unreliability is very small and the network is not too large. We generalize the method in [31] to cover a wider range of applications, in which an estimation problem can be reframed in terms of the hitting time of a given set of states by a continuous-time Markov chain. The estimator is then defined as a function of the sample path of the underlying discrete time chain only, via Conditional Monte Carlo.We prove that the method gives bounded relative error for rare event probability estimation in certain settings. We show how it can be used to estimate the cumulative distribution function, or the density, or some moment of the hitting time. We provide examples for which the method can be applied and we give numerical illustrations.

Another family of models of interest in the group are the highly reliable Markovian systems, where a Markov chain models the evolution of a multicomponent system with failures and repairs of its components. In [27] we explore a new approach in the context of these models, and in the rare event case, called Conditional Monte Carlo with Intermediate Estimations (CMIE). The target are models with complex structures, where it is hard to design a good *importance function* dealing to good Importance Sampling schemes. The paper shows that the method belongs to the variance reduction family, and some examples illustrate its performances. It can be seen as a generalization of the class of splitting simulation procedures.

Finally, in Quasi-Monte Carlo (QMC), we reviewed in [64] the recent development on array-RQMC, a randomized quasi-Monte Carlo method for we had developed estimating the state distribution at each step of a Markov chain with totally ordered (discrete or continuous) state space. It can be used in particular to obtain a low-variance unbiased estimator of the expected total cost up to some random stopping time, when state-dependent costs are paid at each step. In [21], a combination of sequential MC with RQMC to accelerate convergence proposed by Gerber and Chopin is compared with our array-RQMC.

But simulation requires the use of pseudo-random generators. In [45] we provide a review of the state of the art on the design and implementation of random number generators (RNGs) for simulation, on both sequential and parallel computing environments. A general review of pseudo-random and quasi-random number generation is also provided in [73]. A tool for the generation of rank-1 lattice rules is described in [22].

DIVERSE Project-Team

7. New Results

7.1. Results on Software Language Engineering

7.1.1. Modular and Reusable Development of DSLs

Domain-Specific Languages (DSLs) are now developed for a wide variety of domains to address specific concerns in the development of complex systems. When engineering new DSLs, it is likely that previous efforts spent on the development of other languages could be leveraged, especially when their domains overlap. However, legacy DSLs may not fit exactly the end user requirements and thus require further extension, restriction, or specialization. While current language workbenches provide import mechanisms, they usually lack an explicit support for such customizations of imported artifacts. We propose an approach for building DSLs by safely assembling and customizing legacy DSLs artifacts. This approach is based on typing relations that provide a reasoning layer for manipulating DSLs while ensuring type safety. On top of this reasoning layer, we provide an algebra of operators for extending, restricting, and assembling separate DSL artifacts. We implemented the typing relations and algebra into the Melange meta-language [30], [29], [73].

7.1.2. Executable Domain-Specific Modeling Languages (xDSMLs)

Executable Domain-Specific Modeling Languages (xDSMLs) open many possibilities for performing early verification and validation (V&V) of systems. Dynamic V&V approaches rely on execution traces, which represent the evolution of models during their execution. In order to construct traces, generic trace metamodels can be used. Yet, regarding trace manipulations, they lack both efficiency because of their sequential structure, and usability because of their gap to the xDSML. We contributed a generative approach that defines a rich and domain-specific trace metamodel enabling the construction of execution traces for models conforming to a given xDSML [24]. We also contributed a partly generic omniscient debugger supported by generated domain-specific trace management facilities [49].

The emergence of modern concurrent systems calls for xDSMLs where concurrency is of paramount importance. Such xDSMLs are intended to propose constructs with rich concurrency semantics, which allow system designers to precisely define and analyze system behaviors. In [34], we introduce a concurrent executable metamodeling approach, which supports a modular definition of the execution semantics, including the concurrency model, the semantic rules, and a well-defined and expressive communication protocol between them. In [28], we present MoCCML, a dedicated meta-language for formally specifying the concurrency concern within the definition of a DSL. The concurrency constraints can reflect the knowledge in a particular domain, but also the constraints of a particular platform. MoCCML comes with a complete language workbench to help a DSL designer in the definition of the concurrency directly within the concepts of theDSL itself, and a generic workbench to simulate and analyze any model conforming to this DSL. MoCCML is illustrated on the definition of an lightweight extension of SDF (SynchronousData Flow).

7.1.3. Globalization of Domain-Specific Modeling Languages

The development of modern complex software-intensive systems often involves the use of multiple DSMLs that capture different system aspects. Supporting coordinated use of DSMLs leads to what we call the globalization of modeling languages, that is, the use of multiple modeling languages to support coordinated development of diverse aspects of a system.

In a book published in 2015 [66], a number of articles describe the vision and the way globalized DSMLs currently assist integrated DSML support teams working on systems that span many domains and concerns to determine how their work on a particular aspect influences work on other aspects. Globalized DSMLs offer support for communicating relevant information, and for coordinating development activities and associated technologies within and across teams, in addition to providing support for imposing control over development artifacts produced by multiple teams. DSMLs can be used to support socio-technical coordination by providing the means for stakeholders to bridge the gap between how they perceive a problem and its solution, and the programming technologies used to implement a solution. They also support coordination of work across multiple teams. DSMLs developed in an independent manner to meet the specific needs of domain experts have an associated framework that regulates interactions needed to support collaboration and work coordination across different system domains. The book includes [63], [65], [64], [62] with authors from the DIVERSE team.

In [43], we propose a Behavioral Coordination Operator Language (B-COOL) to reify coordination patterns between specific domains by using coordination operators between the Domain-Specific Modeling Languages used in these domains. Those operators are then used to automate the coordination of models conforming to these languages. We illustrate the use of B-COOL with the definition of coordination operators between timed finite state machines and activity diagrams.

The GEMOC Studio (http://gemoc.org/studio) is an eclipse package that contains components for building and composing executable Domain-Specific Modeling Languages (DSMLs). The GEMOC Studio complements Melange to formally define in a modular way the concurrency model of executable DSMLs, and provides analysis and coordination facilities based on the concurrency model. It also integrates all the contributions presented in this document related to model execution, animation, debugging and trace management. The GEMOC studio has been the overall winner of the transformation tool contest 2015 on Model Execution [52].

7.1.4. An analysis of metamodeling practices for MOF and OCL

The definition of a metamodel that precisely captures domain knowledge for effective know-how capitalization is a challenging task. A major obstacle for domain experts who want to build a metamodel is that they must master two radically different languages: an object-oriented, MOF-compliant, modeling language to capture the domain structure and first order logic (the Object Constraint Language) for the definition of well-formedness rules. However, there are no guidelines to assist the conjunct usage of both paradigms, and few tools support it. Consequently, we observe that most metamodels have only an object-oriented domain structure, leading to inaccurate metamodels. In [21], we perform the first empirical study, which analyzes the current state of practice in metamodels that actually use logical expressions to constrain the structure. We analyze 33 metamodels including 995 rules coming from industry, academia and the Object Management Group, to understand how metamodelers articulate both languages. We implement a set of metrics in the OCLMetrics tool to evaluate the complexity of both parts, as well as the coupling between both. We observe that all metamodels tend to have a small, core subset of concepts, which are constrained by most of the rules, in general the rules are loosely coupled to the structure and we identify the set of OCL constructs actually used in rules.

7.1.5. Model Slicers

Among model comprehension tools, model slicers are tools that extract a subset of model elements, for a specific purpose. We propose the Kompren language to model and generate model slicers for any DSL (*e.g.* modeling for software development or for civil engineering) and for different purposes (*e.g.* monitoring and model comprehension). We detail the semantics of the Kompren language and of the model slicer generator. This provides a set of expected properties about the slices that are extracted by the different forms of the slicer [18]. We show how the use of Kompren, a domain-specific language for defining model slicers, can ease the development of such interactive visualization features [19].

In Model Driven Development (MDD), it is important to ensure that a model conforms to the invariants defined in the metamodel. General-purpose rigorous analysis tools that check invariants are likely to perform the analysis over the entire metamodel and model. Since modern day software is exceedingly complex, the

size of the model together with the metamodel can be very large. Consequently, invariant checking can take a very long time.To this end, we introduce model slicing within the invariant checking process, and use a slicing technique to reduce the size of the inputs in order to make invariant checking of large models feasible with existing tools [22], [42].

7.1.6. Bridging the gap between scientific models and engineering models with MDE

The complex problems that computational science addresses are more and more benefiting from the progress of computing facilities (e.g., simulators, librairies, accessible languages). Nevertheless, the actual solutions call for several improvements. Among those, we address in [25] the needs for leveraging on knowledge and expertise by focusing on Domain-Specific Modeling Languages application. In this vision paper we illustrate, through concrete experiments, how the last DSML research help getting closer the problem and implementation spaces.

Various disciplines use models for different purposes. While engineering models, including software engineering models, are often developed to guide the construction of a non- existent system, scientific models, in contrast, are created to better understand a natural phenomenon (i.e., an already existing system). An engineering model may incorporate scientific models to build a system. Both engineering and scientific models have been used to support sustainability, but largely in a loosely-coupled fashion, independently developed and maintained from each other. Due to the inherent complex nature of sustainability that must balance trade-offs between social, environmental, and economic concerns, modeling challenges abound for both the scientific and engineering disciplines. In [72] we propose a vision that synergistically combines engineering and scientific models to enable broader engagement of society for addressing sustainability concerns, informed decision-making based on more accessible scientific models and data, and automated feed-back to the engineering models to support dynamic adaptation of sustainability systems. To support this vision, we identify a number of challenges to be addressed with particular emphasis on the socio-technical benefits of modeling.

As first experiments, we presented at the Inria-Industry meeting 2015 on energy transition and EclipseCon 2015, an approach to develop smart cyber physical systems in charge of managing the production, distribution and consumption of energies (e.g., water, electricity). The main objective is to enable a broader engagement of society, while supporting a more informed decision-making, possibly automatically, on the development and run-time adaptation of sustainability systems (e.g., smart grid, home automation, smart cities). We illustrate this approach through a system that allows farmers to simulate and optimize their water consumption by combining the model of a farming system together with agronomical models (e.g., vegetable and animal lifecycle) and open data (e.g., climate series). To do so, we use Model Driven Engineering (MDE) and Domain Specific Languages (DSL) to develop such systems driven by scientific models that define the context (e.g., environment, social and economy), and model experiencing environments to engage general public and policy makers.

7.2. Results on Variability Modeling and Engineering

7.2.1. Reverse engineering variability

We have developed automated techniques and a comprehensive environment for synthesizing feature models from various kinds of artefacts (e.g. propositional formula, dependency graph, FMs or product comparison matrices). Specifically we have elaborated a support (through ranking lists, clusters, and logical heuristics) for choosing a sound and meaningful hierarchy [93]. We have performed an empirical evaluation on hundreds of feature models, coming from the SPLOT repository and Wikipedia [92]. We have showed that a hybrid approach mixing logical and ontological techniques outperforms state-of-the-art solutions (to appear in Empirical Software Engineering journal in 2015 [20]). We have also considered numerical information and feature *attributes* so that we are now capable of synthesizing attributed feature models from product descriptions [51].

Besides, we have developed techniques for reverse engineering variability in generators and configurators (e.g., video generators) [50]. We have identified new research directions for protecting variability [44] mainly due to the fact reverse engineering techniques (previously presented) are effective .

7.2.2. Product comparison matrices

Product Comparison Matrices (PCMs) constitute a rich source of data for comparing a set of related and competing products over numerous features. PCMs can be seen as a formalism for modeling a family of products, including variability information. Despite their apparent simplicity, PCMs contain heterogeneous, ambiguous, uncontrolled and partial information that hinders their efficient exploitations. We have formalized PCMs through model-based automated techniques and developed additional tooling to support the edition and re-engineering of PCMs [94]. 20 participants used our editor to evaluate our PCM metamodel and automated transformations. The empirical results over 75 PCMs from Wikipedia show that (1) a significant proportion of the formalization of PCMs can be automated: 93.11% of the 30061 cells are correctly formalized; (2) the rest of the formalization can be realized by using the editor and mapping cells to existing concepts of the metamodel. The ASE'2014 paper opens avenues for engaging a community in the mining, re-engineering, edition, and exploitation of PCMs that now abound on the Internet. We have launched an open, collaborative initiative towards this direction https://opencompare.org/

Another axis is the mining of PCMs since (1) the manual elaboration of PCMs has limitations (2) numerous sources of information can be combined and are amenable to PCMs. We have developed MatrixMiner a tool for automatically synthesizing PCMs from a set of product descriptions written in natural language [46]. MatrixMiner is capable of identifying and organizing features and values in a PCM despite the informality and absence of structure in the textual descriptions of products. More information is available online: https://matrix-miner.variability.io/

7.3. Results on Heterogeneous and dynamic software architectures

7.3.1. Resource Monitoring and Reservation in Heterogeneous and dynamic software architectures

Software systems are more pervasive than ever nowadays. Occasionally, applications run on top of resource-constrained devices where efficient resource management is required; hence, they must be capable of coping with such limitations. However, applications require support from the runtime environment to properly deal with resource limitations. This thesis addresses the problem of supporting resource-aware programming in execution environments. In particular, it aims at offering efficient support for collecting data about the consumption of computational resources (e.g., CPU, memory), as well as efficient mechanisms to reserve resources for specific applications. In existing solutions we find two important drawbacks. First, they impose performance overhead on the execution of applications. Second, creating resource management tools for these abstractions is still a daunting task. The outcomes of this work [12] are three contributions:

- An optimistic resource monitoring framework that reduces the cost of collecting resource consumption data.
- A methodology to select components' bindings at deployment time in order to perform resource reservation.
- A language to build customized memory profilers that can be used both during applications' development, and also in a production environment.

7.3.2. Dynamic Reasoning on Heterogeneous and dynamic software architectures

Multi-Objective Evolutionary Algorithms (MOEAs) have been successfully used to optimize various domains such as finance, science, engineering, logistics and software engineering. Nevertheless, MOEAs are still very complex to apply and require detailed knowledge about problem encoding and mutation operators to obtain an effective implementation. Software engineering paradigms such as domain-driven design aim to tackle this complexity by allowing domain experts to focus on domain logic over technical details. Similarly, in order to handle MOEA complexity, we propose an approach, using model-driven software engineering (MDE) techniques, to define fitness functions and mutation operators without MOEA encoding knowledge. Integrated into an open source modelling framework, our approach can significantly simplify development and maintenance of multi-objective optimizations. By leveraging modeling methods, our approach allows reusable

optimizations and seamlessly connects MOEA and MDE paradigms. We evaluate our approach on a cloud case study and show its suitability in terms of i) complexity to implement an MOO problem, ii) complexity to adapt (maintain) this implementation caused by changes in the domain model and/or optimization goals, and iii) show that the efficiency and effectiveness of our approach [56] remains comparable to ad-hoc implementations.

7.3.3. A Precise Metamodel for Open Cloud Computing Interface

Open Cloud Computing Interface (OCCI) proposes one of the first widely accepted, community-based, open standards for managing any kinds of cloud resources. But as it is specified in natural language, OCCI is imprecise, ambiguous, incomplete, and needs a precise definition of its core concepts. Indeed, the OCCI Core Model has conceptual drawbacks: an imprecise semantics of its type classification system, a nonextensible data type system for OCCI attributes, a vague and limited extension concept and the absence of a configuration concept. To tackle these issues, this work proposes a precise metamodel for OCCI. This metamodel defines rigourously the static semantics of the OCCI core concepts, of a precise type classification system, of an extensible data type system, and of both extension and configuration concepts. This metamodel is based on the Eclipse Modeling Framework (EMF), its structure is encoded with Ecore and its static semantics is rigourously defined with Object Constraint Language (OCL). As a consequence, this metamodel provides a concrete language to precisely define and exchange OCCI models. The validation of our metamodel is done on the first worldwide dataset of OCCI extensions already published in the literature, and addressing inter-cloud networking, infrastructure, platform, application, service management, cloud monitoring, and autonomic computing domains, respectively. This validation highlights simplicity, consistency, correctness, completeness, and usefulness of the proposed metamodel[38], [41].

7.3.4. Using Novelty Search Approach and models@runtime for Automatic Testing Environment Setup

In search-based structural testing, metaheuristic search techniques have been frequently used to automate the test data generation. In Genetic Algorithms (GAs) for example, test data are rewarded on the basis of an objective function that represents generally the number of statements or branches covered. However, owing to the wide diversity of possible test data values, it is hard to find the set of test data that can satisfy a specific coverage criterion. In this work, we introduce the use of Novelty Search (NS) algorithm to the test data generation problem based on statement-covered criteria. We believe that such approach to test data generation is attractive because it allows the exploration of the huge space of test data within the input domain. In this approach, we seek to explore the search space without regard to any objectives. In fact, instead of having a fitness-based selection, we select test cases based on a novelty score showing how different they are compared to all other solutions evaluated so far [47], [48]. We also create an architecture generation framework for setup testing environment for a distributed and heterogeneous service.

7.3.5. Using Models@Run.time to embed an Energetic Cloud Simulator in a MAPE-K Loop

Due to high electricity consumption in the Cloud datacenters, providers aim at maximizing energy efficiency through VM consolidation, accurate resource allocation or adjusting VM usage. More generally, the provider attempts to optimize resource utilization. However, while minimizing expenses, the Cloud operator still needs to conform to SLA constraints negotiated with customers (such as latency, downtime, affinity, placement, response time or duplication). Consequently, optimizing a Cloud configuration is a multi-objective problem. As a nontrivial multi-objective optimization problem, there does not exist a single solution that simultaneously optimizes each objective. There exists a (possibly infinite) number of Pareto optimal solutions. Evolutionary algorithms are popular approaches for generating Pareto optimal solutions to a multi-objective optimization problem. Most of these solutions use a fitness function to assess the quality of the candidates. However, regarding the energy consumption estimation, the fitness function can be approximative and lead to some imprecisions compared to the real observed data. This work presents a system that uses a genetic algorithm to optimize Cloud energy consumption and machine learning techniques to improve the fitness function regarding a real distributed cluster of server. We have carried out experiments on the OpenStack platform to validate our solution. This experimentation shows that the machine learning produces an accurate energy model, predicting precise values for the simulation [124][40].

7.4. Results on Diverse Implementations for Resilience

Diversity is acknowledged as a crucial element for resilience, sustainability and increased wealth in many domains such as sociology, economy and ecology. Yet, despite the large body of theoretical and experimental science that emphasizes the need to conserve high levels of diversity in complex systems, the limited amount of diversity in software-intensive systems is a major issue. This is particularly critical as these systems integrate multiple concerns, are connected to the physical world through multiple sensors, run eternally and are open to other services and to users. Here we present our latest observational and technical results about (i) new approaches to increase diversity in software systems, and (ii) software testing to assess the validity of software.

7.4.1. Software diversification

Early experiments with software diversity in the mid 1970's investigated N-version programming and recovery blocks to increase the reliability of embedded systems. Four decades later, the literature about software diversity has expanded in multiple directions: goals (fault-tolerance, security, software engineering); means (managed or automated diversity) and analytical studies (quantification of diversity and its impact). We contribute to the field of software diversity with the very first literature survey that adopts an inclusive vision of the area, with an emphasis on the most recent advances in the field. This survey includes classical work about design and data diversity for fault tolerance, as well as the cybersecurity literature that investigates randomization at different system levels. It broadens this standard scope of diversity, to include the study and exploitation of natural diversity and the management of diverse software products [17].

We also contribute to software diversity with novel techniques and methods. The interdisciplinary investigations within the DIVERSIFY project have led to the definition of novel principles for open-ended evolution in software systems. The main intuition is that software should have the ability to spontaneously and continuously evolve without waiting for specific environmental conditions. Our proposal analogizes the software consumer / provider network, which can be found in any types of distributed systems, to a bipartite ecological graph. This analogy provides the foundations for the design of an individual-based simulator used to experiment with decentralized adaptation strategies for providers and consumers. The initial model of a software network is tuned according to observations gathered from real-world software networks. The key insights about our experiments are that, 1) we can successfully model software systems as an ALife system, and 2) we succeed in emerging a global property from local decisions: when consumers and providers adapt with local decision strategies, the global robustness of the network increases. We show that these results hold with different initial situations, different scales and different topological constraints on the network [55]. In order to move towards the open-ended evolution of actual systems, we also developed a novel tool for the runtime modification of Java programs, as an extension to the JVM [60].

Our second contribution to the filed of software diversity consists in experimenting its application in different fields. First, we have proposed a novel approach to exploit software diversity at multiple granularity levels simultaneously [15]. The main idea is to reconcile two aspects of the massive software reuse in web applications: on the one hand, reuse and modularity favor much writing the next killer application; on the other hand, reuse and modularity facilitates much the next massive BOBE attack. We demonstrate the feasiblity of diversifying web applications at multiple levels, mitigating the risks of reuse.

The second application of automatic software diversification for Java programs aimed at answering the following question: which product line operators, applied to which program elements , can synthesize variants of programs that are incorrect , correct or perhaps even conforming to test suites? We implement source code transformations, based on the derivation operators of the Common Variability Language. We automatically synthesize more than 370,000 program variants from a set of 8 real large Java projects (up to 85,000 lines of code), obtaining an extensive panorama of the sanity of the operations [68].

The third application of software diversification is against browser fingerprinting. Browser fingerprint tracking relies on the following mechanisms: web browsers allow remote servers to discover sufficient information about a user's platform to create a digital fingerprint that uniquely identifies the platform. We argue that fingerprint uniqueness and stability are the key threats to browser fingerprint tracking, and we aim at breaking fingerprint stability over time, by exploiting software diversity and automatic reconfiguration. We leverage

virtualization and modular software architectures to automatically assemble and reconfigure a user's software components at multiple levels. We operate on the operating system, the browser, the lists of fonts and plugins. This work is the first application of software reconfiguration to build a moving target defense against browser fingerprint tracking. We have developed a prototype called *Blink* to experiment the effectiveness of our approach at randomizing fingerprints [33].

7.4.2. Software testing

Our work in the area of software testing focuses on tailoring the testing tools (analysis, generation, oracle, etc.) to specific domains. This allows us to consider domain specific knowledge (e.g., architectural patterns for GUI implementation) in order to increase the relevance and the efficiency of testing. The main results of this year are about testing GUIs and model transformations.

Graphical user interfaces (GUIs) are integral parts of software systems that require interactions from their users. Software testers have paid special attention to GUI testing in the last decade, and have devised techniques that are effective in finding several kinds of GUI errors. However, the introduction of new types of interactions in GUIs presents new kinds of errors that are not targeted by current testing techniques. We believe that to advance GUI testing, the community needs a comprehensive and high level GUI fault model, which incorporates all types of interactions. In this work, we first propose a GUI fault model designed to identify and classify GUI faults [37]. We then studied the impact of the new types of interactions in GUIs on their testing process. We show that the current GUI model-based testing approaches have limits when applied to test such new advanced GUIs [36].

Specifying a model transformation is challenging as it must be able to give a meaningful output for any input model in a possibly infinite modeling domain. Transformation preconditions constrain the input domain by rejecting input models that are not meant to be transformed by a model transformation. In our latest work [39], we present a systematic approach to discover such preconditions when it is hard for a human developer to foresee complex graphs of objects that are not meant to be transformed. The approach is based on systematically generating a finite number of test models using our tool, PRAMANA to first cover the input domain based on input domain partitioning. Tracing a transformation's execution reveals why some preconditions are missing. Using a benchmark transformation from simplified UML class diagram models to RDBMS models we discover new preconditions that were not initially specified.

We also initiated a new line of research in order to invesitgate Novelty Search (NS) for the automatic generation of test data. This allows the exploration of the huge space of test data within the input domain. In this approach, we select test cases based on a novelty score showing how different they are compared to all other solutions evaluated so far [47].

In Model Driven Engineering (MDE), models are first-class citizens, and model transformation is MDE's "heart and soul". Since model transformations are executed for a family of (conforming) models, their validity becomes a crucial issue. In [16] we proposes to explore the question of the formal verification of model transformation properties through a tridimensional approach: the transformation involved, the properties of interest addressed, and the formal verification techniques used to establish the properties. This work is intended for a double audience. For newcomers, it provides a tutorial introduction to the field of formal verification of model transformations. For readers more familiar with formal methods and model transformations, it proposes a literature review (although not systematic) of the contributions of the field. Overall, this work allows to better understand the evolution, trends and current practice in the domain of model transformation verification. This work opens an interesting research line for building an engineering of model transformation verification guided by the notion of model transformation intent.

KERDATA Project-Team

7. New Results

7.1. Efficient data management for hybrid and multi-site clouds

7.1.1. JetStream: enabling high-throughput live event streaming on multi-site clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Scientific and commercial applications operate nowadays on tens of cloud datacenters around the globe, following similar patterns: they aggregate monitoring or sensor data, assess the QoS or run global data mining queries based on inter-site event stream processing. Enabling fast data transfers across geographically distributed sites allows such applications to manage the continuous streams of events in real time and quickly react to changes. However, traditional event processing engines often consider data resources as second-class citizens and support access to data only as a side-effect of computation (i.e. they are not concerned by the transfer of events from their source to the processing site). This is an efficient approach as long as the processing is executed in a single cluster where nodes are interconnected by low latency networks. In a distributed environment, consisting of multiple datacenters, with orders of magnitude differences in capabilities and connected by a WAN, this will undoubtedly lead to significant latency and performance variations.

This is namely the challenge we addressed this year by proposing JetStream [15], a high performance batch-based streaming middleware for efficient transfers of events between cloud datacenters. JetStream is able to self-adapt to the streaming conditions by modeling and monitoring a set of context parameters. It further aggregates the available bandwidth by enabling multi-route streaming across cloud sites, while at the same time optimizing resource utilization and increasing cost efficiency. The prototype was validated on tens of nodes from US and Europe datacenters of the Windows Azure cloud with synthetic benchmarks and a real-life application monitoring the ALICE experiment at CERN. The results show a $3\times$ increase of the transfer rate using the adaptive multi-route streaming, compared to state of the art solutions.

7.1.2. Multi-site metadata management for geographically distributed cloud workflows

Participants: Luis Eduardo Pineda Morales, Alexandru Costan, Gabriel Antoniu.

With their globally distributed datacenters, clouds now provide an opportunity to run complex large-scale applications on dynamically provisioned, networked and federated infrastructures. However, there is a lack of tools supporting data-intensive applications (e.g. scientific workflows) on virtualized IaaS or PaaS systems across geographically distributed sites. As a relevant example, data-intensive scientific workflows struggle in leveraging such distributed cloud platforms. For instance, scientific workflows which handle many small files can easily saturate state-of-the-art distributed filesystems based on centralized metadata servers (e.g., HDFS, PVFS).

In [22], we explore several alternative design strategies to efficiently support the execution of existing workflow engines across multi-site clouds, by reducing the cost of metadata operations. These strategies leverage workflow semantics in a 2-level metadata partitioning hierarchy that combines distribution and replication. The system was validated on the Microsoft Azure cloud across 4 EU and US datacenters. The experiments were conducted on 128 nodes using synthetic benchmarks and real-life applications. We observe as much as 28% gain in execution time for a parallel, geo-distributed real-world application (Montage) and up to 50% for a metadata-intensive synthetic benchmark, compared to a baseline centralized configuration.

7.1.3. Understanding the performance of Big Data platforms in hybrid and multi-site clouds

Participants: Roxana-Ioana Roman, Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu.

Recently, hybrid multi-site big data analytics (that combines on-premise with off-premise resources) has gained increasing popularity as a tool to process large amounts of data on-demand, without additional capital investment to increase the size of a single datacenter. However, making the most out of hybrid setups for big data analytics is challenging because on-premise resources can communicate with off-premise resources at significantly lower throughput and higher latency. Understanding the impact of this aspect is not trivial, especially in the context of modern big data analytics frameworks that introduce complex communication patterns and are optimized to overlap communication with computation in order to hide data transfer latencies. This year we started to work on a study that aims to identify and explain this impact in relationship to the known behavior on a single cloud.

A first step towards this goal consisted of analysing a representative big data workload on a hybrid Spark setup [24]. Unlike previous experience that emphasized low end-impact of network communications in Spark, we found significant overhead in the shuffle phase when the bandwidth between the on-premise and off-premise resources is sufficiently small. We plan to continue this study by investigating additional parameters at a finer grain and adding new platforms, like Apache Flink.

7.2. Optimizing Map-Reduce

7.2.1. Chronos: failure-aware scheduling in shared Hadoop clusters

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Hadoop emerged as the de facto state-of-the-art system for MapReduce-based data analytics. The reliability of Hadoop systems depends in part on how well they handle failures. Currently, Hadoop handles machine failures by re-executing all the tasks of the failed machines (i.e., executing recovery tasks). Unfortunately, this elegant solution is entirely entrusted to the core of Hadoop and hidden from Hadoop schedulers. The unawareness of failures therefore may prevent Hadoop schedulers from operating correctly towards meeting their objectives (e.g., fairness, job priority) and can significantly impact the performance of MapReduce applications.

In [23], we propose Chronos, a failure-aware scheduling strategy that enables an early yet smart action for fast failure recovery while operating within a specific scheduler objective. Chronos takes an early action rather than waiting an uncertain amount of time to get a free slot (thanks to our preemption technique). Chronos embraces a smart selection algorithm that returns a list of tasks that need to be preempted in order to free the necessary slots to launch recovery tasks immediately. This selection considers three criteria: the progress scores of running tasks, the scheduling objectives, and the recovery tasks input data locations. In order to make room for recovery tasks rather than waiting an uncertain amount of time, a natural solution is to kill running tasks in order to create free slots. Although killing tasks can free the slots easily, it wastes the work performed by the killed tasks. Therefore, we present the design and implementation of a novel work-conserving preemption technique that allows pausing and resuming both map and reduce tasks without resource wasting and with little overhead.

We demonstrate the utility of Chronos by combining it with two state-of-the-art Hadoop schedulers: Fifo and Fair schedulers. The experimental results show that Chronos achieves almost optimal data locality for the recovery tasks and reduces the job completion times by up to 55% over state-of-the-art schedulers. Moreover, Chronos recovers to a correct scheduling behavior after failure detection within only a couple of seconds.

7.2.2. On the usability of shortest remaining time first policy in shared Hadoop clusters

Participants: Nathanaël Cheriere, Shadi Ibrahim.

A practical problem facing the Hadoop community is how to reduce job makespans by reducing job waiting times and execution times. Previous Hadoop schedulers have focused on improving job execution times, by improving data locality but not considering job waiting times. Even worse, enforcing data locality according to the job input sizes can be inefficient: it can lead to long waiting times for small yet short jobs when sharing the cluster with jobs with smaller input sizes but higher execution complexity.

We have introduced hSRTF [16], an adaption of the well-known Shortest Remaining Time First scheduler (i.e., SRTF) in shared Hadoop clusters. hSRTF embraces a simple model to estimate the remaining time of a job and a preemption primitive (i.e., kill) to free the resources when needed. We have implemented hSRTF and performed extensive evaluations with Hadoop on the Grid'5000 testbed. The results show that hSRTF can significantly reduce the waiting times of small jobs and therefore improves their make-spans, but at the cost of a relatively small increase in the make-spans of large jobs. For instance, a time-based proportional share mode of hSRTF (i.e., hSRTF-Pr) speeds up small jobs by (on average) 45% and 26% while introducing a performance degradation for large jobs by (on average) 10% and 0.2% compared to Fifo and Fair schedulers, respectively.

7.2.3. A Performance evaluation of Hadoop's schedulers under failures

Participants: Shadi Ibrahim, Gabriel Antoniu.

Recently, Hadoop has not only been used for running single batch jobs but it has also been optimized to simultaneously support the execution of multiple jobs belonging to multiple concurrent users. Several schedulers (i.e., Fifo, Fair, and Capacity schedulers) have been proposed to optimize locality executions of tasks but do not consider failures, although, evidence in the literature shows that faults do occur and can probably result in performance problems.

In [19], we have designed a set of experiments to evaluate the performance of Hadoop under failure when applying several schedulers (i.e., explore the conflict between job scheduling, exposing locality executions, and failures). Our results reveal several drawbacks of current Hadoop's mechanism in prioritizing failed tasks. By trying to launch failed tasks as soon as possible regardless of locality, it significantly increases the execution time of jobs with failed tasks, due to two reasons: 1) available resources might not be freed up as quickly as expected and 2) failed tasks might be re-executed on machines with no data on it, introducing extra cost for data transferring through network, which is normally the most scarce resource in today's datacenters.

Our preliminary study with Hadoop not only helps us to understand the interplay between fault-tolerance and job scheduling, but also offers useful insights into optimizing the current schedulers to be more efficient in case of failures.

7.2.4. Kvasir: empowering Hadoop with knowledge

Participants: Nathanaël Cheriere, Shadi Ibrahim.

Most of Hadoop schedulers are based on homogeneity hypotheses about the jobs and the nodes and therefore strongly rely on the location of the input data when scheduling tasks. However, our study revealed that Hadoop is a highly dynamic environment (e.g., variation in task duration within a job and across different jobs). Even worse, clouds are multi-tenant environments which in turn introduce more heterogeneity and dynamicity in Hadoop clusters. As a result, relying on static knowledge (i.e. data location) may lead to wrong scheduling decisions.

We have developed a new scheduling framework for Hadoop, named Kvasir. Kvasir aims to provide an up-to-date knowledge that reflects the dynamicity of the environment while being light-weight and performance-oriented. The utility of Kvasir is demonstrated by the implementation of several schedulers including Fifo, Fair, and SRTF schedulers.

7.3. Energy-aware data management in clouds and HPC

7.3.1. On understanding the energy impact of speculative execution in Hadoop

Participants: Tien Dat Phan, Shadi Ibrahim, Gabriel Antoniu, Luc Bougé.

Hadoop emerged as an important system for large-scale data analysis. Speculative execution is a key feature in Hadoop that is extensively leveraged in clouds: it is used to mask slow tasks (i.e., stragglers) — resulted from resource contention and heterogeneity in clouds — by launching speculative task copies on other machines. However, speculative execution is not cost-free and may result in performance degradation and extra resource and energy consumption. While prior literature has been dedicated to improving stragglers detection to cope with the inevitable heterogeneity in clouds, little work is focusing on understanding the implications of speculative execution on the performance and energy consumption in Hadoop cluster.

In [21], we have designed a set of experiments to evaluate the impact of speculative execution on the performance and energy consumption of Hadoop in homogeneous and heterogeneous environments. Our studies reveal that speculative execution may sometimes reduce, sometimes increase the energy consumption of Hadoop clusters. This strongly depends on the reduction in the execution time of MapReduce applications and on the extra power consumption introduced by speculative execution. Moreover, we show that the extra power consumption varies among applications and is contributed to by three main factors: the duration of speculative tasks, the idle time, and the allocation of speculative tasks. To the best of our knowledge, our work provides the first deep look into the energy efficiency of speculative execution in Hadoop.

7.3.2. On the energy footprint of I/O management in Exascale HPC systems

Participants: Orçun Yildiz, Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

The advent of unprecedentedly scalable yet energy hungry Exascale supercomputers poses a major challenge in sustaining a high performance-per-watt ratio. With I/O management acquiring a crucial role in supporting scientific simulations, various I/O management approaches have been proposed to achieve high performance and scalability. However, the details of how these approaches affect energy consumption have not been studied yet.

Therefore, we have explored how much energy a supercomputer consumes while running scientific simulations when adopting various I/O management approaches. In particular, we closely examined three radically different I/O schemes including time partitioning, dedicated cores, and dedicated nodes. To do so, we implemented the three approaches within the Damaris I/O middleware and performed extensive experiments with one of the target HPC applications of the Blue Waters sustained-petaflop supercomputer project: the CM1 atmospheric model.

Our experimental results obtained on the French Grid'5000 platform highlighted the differences among these three approaches and illustrate in which way various configurations of the application and of the system can impact performance and energy consumption. Considering that choosing the most energy-efficient approach for a particular simulation on a particular machine can be a daunting task, we provided a model to estimate the energy consumption of a simulation under different I/O approaches. Our proposed model gives hints to pre-select the most energy-efficient I/O approach for a particular simulation on a particular HPC system and therefore provides a step towards energy-efficient HPC simulations in Exascale systems.

We validated the accuracy of our proposed model using a real-life HPC application (CM1) and two different clusters provisioned on the Grid'5000 testbed. The estimated energy consumptions are within 5.7% of the measured ones for all I/O approaches.

7.3.3. Exploring energy-consistency trade-offs in cloud storage systems and beyond

Participants: Mohammed-Yacine Taleb, Shadi Ibrahim, Gabriel Antoniu, Luc Bougé.

Apache Cassandra is an open-source cloud storage system that offers multiple types of operation-level consistency including eventual consistency with multiple levels of guarantees and strong consistency. It is being used by many datacenter applications (e.g., Facebook and AppScale). Most existing research efforts have been dedicated to exploring trade-offs such as: consistency vs. performance, consistency vs. latency and consistency vs. monetary cost. In contrast, a little work is focusing on the consistency vs. energy trade-off. As power bills have become a substantial part of the monetary cost for operating a datacenter, we aim to provide a clearer understanding of the interplay between consistency and energy consumption.

In [17], a series of experiments have been conducted to explore the implication of different factors on the energy consumption in Cassandra. Our experiments have revealed a noticeable variation in the energy consumption depending on the consistency level. Furthermore, for a given consistency level, the energy consumption of Cassandra varies with the access pattern and the load exhibited by the application. This further analysis indicated that the uneven distribution of the load amongst different nodes also impacts the energy consumption in Cassandra. Finally, we experimentally compared the impact of four storage configuration and data partitioning policies on the energy consumption in Cassandra: interestingly, we achieve 23% energy saving when assigning 50% of the nodes to the hot pool for the applications with moderate ratio of reads and writes, while applying eventual (quorum) consistency.

This study points to opportunities for future research on consistency-energy trade-offs and offers useful insight into designing energy-efficient techniques for cloud storage systems. This work was done in collaboration with Houssem-Eddine Chihoub (LIG lab, Grenoble) and María Pérez (UPM, Madrid).

Recently, we have been looking at in-memory storage systems. In particular, we are investigating the current replication schemes, data placement strategies and consistency models which are used in in-memory storage systems. Next, an empirical study will be performed to analyze the potential impact of the aforementioned issues on energy consumption. At this point, we are working with RAMCloud.

7.3.4. Governing energy consumption in Hadoop through CPU frequency scaling: an analysis

Participants: Tien Dat Phan, Shadi Ibrahim, Gabriel Antoniu.

In [12], we studied the impact of different existing DVFS (*Dynamic Voltage and Frequency Scaling*) governors (i.e., performance, powersave, on-demand, conservative and userspace) on Hadoop's performance and power efficiency. Interestingly, our experimental results reported not only a noticeable variation of the power consumption and performance with different applications and under different governors, but also demonstrate the opportunity to achieve a better tradeoff between performance and power consumption.

The primary contributions of this work are as follows: (1) it provides an overview of the state-of-the-art techniques for energy-efficiency in Hadoop; (2) it discusses and demonstrates the need for exploiting DVFS techniques for energy reduction in Hadoop; (3) it experimentally demonstrates that MapReduce applications experience variations in performance and power consumption under different CPU frequencies and also under different governors. A micro-analysis section is provided to explain this variation and its cause; (4) it illustrates in practice how the behavior of different governors influences the execution of MapReduce applications and how it shapes the performance of the entire cluster; (5) it also brings out the differences between these governors and CPU frequencies and shows that they are not only sub-optimal for different applications but also sub-optimal for different stages of MapReduce execution; (6) it demonstrates that achieving better energy efficiency in Hadoop cannot be done simply by tuning the governor parameters, nor through a naive coarse-grained tuning of the CPU frequencies or the governors according to the running phase (i.e., map phase or reduce phase).

7.4. Scalable I/Os: visualization and processing

7.4.1. Modeling and predicting I/O patterns of large-scale simulations

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

The increasing gap between the computation performance of post-petascale machines and the performance of their I/O subsystem has motivated many I/O optimizations including prefetching, caching, and scheduling. In order to further improve these techniques, modeling and predicting spatial and temporal I/O patterns of HPC applications as they run has become crucial. Our work in this context focuses on Omnisc'IO, an approach that builds a grammar-based model of the I/O behavior of HPC applications and uses it to predict when future I/O operations will occur, and where and how much data will be accessed. To infer grammars, Omnisc'IO is based on StarSequitur, a novel algorithm extending Nevill-Manning's Sequitur algorithm [11]. Omnisc'IO is transparently integrated into the POSIX and MPI I/O stacks and does not require any modification in applications or higher-level I/O libraries. It works without any prior knowledge of the application and converges to accurate predictions of any N future I/O operations within a couple of iterations. Its implementation is efficient in both computation time and memory footprint.

7.4.2. In situ analysis and visualization workflows

Participants: Matthieu Dorier, Lokman Rahmani, Gabriel Antoniu.

In situ visualization has been proposed in the past few years to couple running simulations with parallel visualization and analysis tools. While many parallel visualization tools now provide in situ visualization capabilities, the trend has been to feed such tools with what previously was large amounts of unprocessed output data and let them render everything at the highest possible resolution. This leads to an increased run time of simulations that still have to complete within a fixed-length job allocation. In this work, we tackle the challenge of enabling in situ visualization under performance constraints. Our approach shuffles data across processes according to its content and filters out part of it in order to feed a visualization pipeline with only a reorganized subset of the data produced by the simulation. Our framework monitors its own performance and reconfigures itself dynamically to achieve the best possible visual fidelity within predefined performance constraints. Experiments on the Blue Waters supercomputer with the CM1 simulation show that our approach enables a $5\times$ speedup and is able to meet performance constraints.

7.5. Scalable storage for data-intensive applications

7.5.1. OverFlow: multi-site aware Big Data management for scientific workflows on clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

The global deployment of cloud datacenters is enabling large-scale scientific workflows to improve performance and deliver fast responses. This unprecedented geographical distribution of the computation is doubled by an increase in the scale of the data handled by such applications, bringing new challenges related to the efficient data management across sites. High throughput, low latencies or cost-related trade-offs are just a few concerns for both cloud providers and users when it comes to handling data across datacenters. Existing solutions are limited to cloud-provided storage, which offers low performance based on rigid cost schemes. In turn, workflow engines need to improvise substitutes, achieving performance at the cost of complex system configurations, maintenance overheads, reduced reliability and reusability.

In [14], we introduced OverFlow, a uniform data-management system for scientific workflows running across geographically distributed sites, aiming to reap economic benefits from this geo-diversity. Our solution is environment-aware, as it monitors and models the global cloud infrastructure, offering high and predictable data-handling performance for transfer cost and time, within and across sites. OverFlow proposes a set of pluggable services, grouped in a data-scientist cloud kit. They provide the applications with the possibility to monitor the underlying infrastructure, to exploit smart data compression, deduplication and geo-replication, to evaluate data-management costs, to set a tradeoff between money and time, and optimize the transfer strategy accordingly. The system was validated on the Microsoft Azure cloud across its 6 EU and US datacenters. The experiments were conducted on hundreds of nodes using synthetic benchmarks and real-life bio-informatics applications (A-Brain, BLAST). The results show that our system is able to model the cloud performance accurately and to leverage this for efficient data dissemination, being able to reduce the monetary costs and transfer time by up to 3 times.

7.5.2. Efficient transactional storage for data-intensive applications

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

As the computational power used by large-scale applications increases, the amount of data they need to manipulate tends to increase as well. A wide range of such applications require robust and flexible storage support for atomic, durable and concurrent transactions. Historically, databases have provided the *de facto* solution to transactional data management, but they have forced applications to drop control over data layout and access mechanisms, while remaining unable to meet the scale requirements of Big Data. More recently, key-value stores have been introduced to address these issues. However, this solution does not provide transactions, or only restricted transaction support, constraining users to carefully coordinate access to data in order to avoid race conditions, partial writes, overwrites, and other hard problems that cause erratic behavior.

We argue that there is a gap between existing storage solutions and application requirements that limits the design of transaction-oriented data-intensive applications. We have started working on a prototype of a massively parallel distributed transactional blob storage system, aiming to fill this gap.

MYRIADS Project-Team

7. New Results

7.1. Cloud Resource Management

Participants: Ancuta Iordache, Christine Morin, Ghada Moualla, Guillaume Pierre, Matthieu Simonin, Lodewijck Vogelzang.

7.1.1. Application Performance Modeling in Heterogeneous Cloud Environments

Participants: Ancuta Iordache, Lodewijck Vogelzang, Guillaume Pierre.

Heterogeneous cloud platforms offer many possibilities for applications to make fine-grained choices over the types of resources they execute on. This opens for example opportunities for fine-grained 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. We significantly improved the speed at which the system can characterize the performance of an arbitrary application. A first publication on this topic has been published [26], and a second one is in preparation.

7.1.2. Heterogeneous Resource Management

Participants: Ancuta Iordache, Guillaume Pierre.

During her internship at Maxeler Technologies, Ancuta Iordache developed an original technique for virtualizing FPGAs such that they can be used as high-performance computing devices in cloud infrastructures. Virtual FPGAs can be accessed remotely by any VM in the system. They can span multiple physical FPGA, they are elastic, and they can also be shared between multiple tenants. A publication on this topic is currently under evaluation.

7.1.3. Self-adaptatable Hadoop Virtual Clusters

Participants: Christine Morin, Ghada Moualla, Matthieu Simonin.

In the context of Ghana Moualla's Master internship, we designed the Elastic MapReduce Adaptation (EMRA) system to execute Hadoop MapReduce applications with user-defined deadlines in cloud virtual clusters. EMRA integrates an algorithm to automatically adapt the Hadoop cluster size at runtime in order to meet user-defined deadlines. We proposed an automatic scaling algorithm, which monitors the progress of the Map phase of the application during its execution and estimate if the user-defined deadline can be met. If the current allocated resources are not sufficient to meet the deadline, more resources are provisioned. The adaptation service comprises of three main components: (i) a monitor to check the progress of the running application, (ii) an estimator to predict the time needed to complete the application based on its current progress ; (iii) a controller to adapt the size of the virtual cluster by adding virtual machines as needed. The controller takes into account the start-up overhead of the new virtual machines and the time needed for the VM to fetch their input data from the original nodes over the network in order to start their map tasks. We implemented a prototype of the EMRA system in the context of Sahara, an environment for managing Hadoop virtual clusters on top of OpenStack IaaS clouds. We experimented the EMRA system on Grid'5000 with traditional MapReduce benchmarks. We evaluated the relative error of the estimator, the cost for scaling up or down a virtual cluster and showed that the proposed adaptation algorithm allows user-defined deadlines to be met.

7.2. Distributed Cloud Computing

Participants: Teodor Crivat, Yvon Jégou, Vlad Mirel, Christine Morin, Anne-Cécile Orgerie, Edouard Outin, Nikolaos Parlavantzas, Jean-Louis Pazat, Guillaume Pierre, Aboozar Rajabi, Carlos Ruiz Diaz, Arnab Sinha, Genc Tato, Cédric Tedeschi.

7.2.1. A multi-objective adaptation system for the management of a Distributed Cloud

Participants: Yvon Jégou, Edouard Outin, Jean-Louis Pazat.

In this project, we consider a "Distributed Cloud" made of multiple data/computing centers interconnected by a high speed network and belonging to the same administration domain. Moreover, in the Cloud organization targeted here, the network capabilities can be dynamically configured in order to guarantee QoS for streaming or to negotiate bandwidth for example.

As a first step, we are focusing on a single centralised Cloud.

Due to the dynamic capabilities of the Clouds, often referred to as elasticity, there is a strong need to dynamically adapt both platforms and applications to users needs and environmental constraints such as electrical power consumption.

We address the management of a Cloud in order to consider both optimization for energy consumption and for users' QoS needs. The objectives of this optimization will be negotiated as contracts on Service Level Agreement (SLA). A special emphasis will be put on the distributed aspect of the platform and include both servers and network adaptation capabilities.

The design of the system relies on self-* techniques and on adaptation mechanisms at any level (from IaaS to SaaS). The MAPE-k framework (Monitor-Analysis-Planning-Execution based on knowledge) is used for the implementation of the system. The technical developments are based on the Openstack framework.

We have implemented a system that uses a genetic algorithm to optimize Cloud energy consumption and machine learning techniques to improve the fitness function regarding a real distributed cluster of servers. We have carried out experiments on the OpenStack platform to validate our solution. This experimentation shows that the machine learning produces an accurate energy model, predicting precise values for the simulation.

We are currently refining this model and comparing it to real measurements on the platform.

This work is done in cooperation with the DIVERSE team and in cooperation with Orange under the umbrella of the B-COM Technology Research Center.

7.2.2. Dynamic reconfiguration for multi-cloud applications

Participants: Nikolaos Parlavantzas, Aboozar Rajabi, Carlos Ruiz Diaz, Arnab Sinha.

In the context of the PaaSage European project, we are working on model-based, continuous self-optimization of multi-cloud applications. In particular, we are developing a dynamic adaptation system, capable of transforming the currently running application configuration into a target configuration in a cost-effective and safe manner. In 2015, we have improved and extended the Adapter prototype [45]. The system now fully supports dynamic configuration, including detecting changes, generating reconfiguration plans, validating plans based on a cost-benefit calculation, and executing plans in parallel, improving adaptation performance. Moreover, we have performed initial investigations on the use of PaaSage for supporting Internet of Things (IoT) applications [27]. Finally, in the context of Carlos Ruiz's stay, we are defining a model for managing the configuration of cloud applications and environments. This model is based on feature modeling and the derived configurations are mapped to PaaSage models.

7.2.3. Towards a distributed cloud inside the backbone

Participants: Christine Morin, Anne-Cécile Orgerie, Genc Tato, Cédric Tedeschi.

The DISCOVERY proposal officially started at the end of 2015. It is an Inria Project Lab (IPL) led by Adrien Lebre from the ASCOLA team, and currently on leave at Inria. It aims at designing a distributed cloud, leveraging the resources we can find in the network backbone. [0] In practice, this work is intended to get integrated within the OpenStack softwarehttps://www.openstack.org/ so as to decentralize its whole architecture.

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

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. In this framework, the PhD thesis of Genc Tato started in December 2015. It aims at developing locality mechanisms at the data management layer.

We have also started an energy/cost-benefit analysis of a decentralized Cloud infrastructure like the one proposed within Discovery. This work is conducted by Anthony Simonet, a post-doctoral researcher on an Inria contract for the Discovery IPL and co-supervised by Adrien Lebre from the ASCOLA team and Anne-Cécile Orgerie from Myriads team.

7.2.4. Mobile edge cloud computing with ConPaaS

Participants: Teodor Crivat, Vlad Mirel, Guillaume Pierre.

Interactive multi-user applications usually rely on intermediate cloud servers to mediate the inter-user interaction. However, current mobile networks exhibit network latencies in the order of 50-150 ms between the device and any cloud. Such latencies make it impossible to create smooth interactions with the end user. To enable an "instantaneous" feeling, augmented reality applications require that end-to-end latencies should remain below 20 ms.

To address these issues, we extended ConPaaS to support the deployment of cloud applications in a distributed set of Raspberry Pi machines. The motivation is to reduce the latency compared to a traditional deployment where the backend is located in an external cloud: instead of reaching the cloud through a wide-area network, in this setup each cloud node is also equipped with a wifi hotspot which allows local users to access it directly.

7.2.5. Fog Computing

Participant: Jean-Louis Pazat.

The concept of "Fog Computing" is currently developed on the idea of hosting instances of services not on centralized datacenters (i.e. the "Cloud"), but on a highly distributed infrastructure: the Internet Edge (i.e. the "Fog"). This infrastructure consists in geographically distributed computing resources with relatively small capabilities. Compared with datacenters, a "Fog" infrastructure is able to offer to Service Providers a shorter distance from the service to the user but with the same flexibility of software deployment and management.

This work focus on the problem of resource allocation in such infrastructure when considering services in the area of Internet of Things, Social Networks or Online Gaming. For such use-cases, service-to-user latency is a critical parameter for the quality of experience. Optimizing such a parameter is an objective for the platform built on top of the Fog Infrastructure that will be dedicated to the deployment of the considered service. In order to achieve such a goal, the platform needs to select some strategies for the allocation of network and computing resources, based on the initial requirements for service distribution.

We are designing a prototype based on micro services and we are considering low overhead virtualization systems using containers. This prototype is intended to run inside an Internet Box or inside a LAN disk server at user's home. The whole system will be intended to be used very small or medium size user communities willing to share devices and data. The main characteristics of the system will be reliable distributed storage and distributed execution of services.

This work is part of Bruno Stevant's PhD thesis, which began in December 2014. It is done in cooperation with the REOP team, Institut Mines telecom/IRISA.

7.3. Cloud Security

Participants: Anna Giannakou, Christine Morin, Jean-Louis Pazat, Louis Rilling, Amir Teshome Wonjiga.

7.3.1. Security Monitoring of Clouds

Participants: Anna Giannakou, Christine Morin, Jean-Louis Pazat, Louis Rilling, Amir Teshome Wonjiga.

We aim at making security monitoring a dependable service for IaaS cloud customers. To this end, we study three topics:

- defining relevant SLA terms for security monitoring,
- enforcing and evaluating SLA terms,
- making the SLA terms enforcement mechanisms self-adaptable to cope with the dynamic nature of clouds.

The considered enforcement and evaluation mechanisms should have a minimal impact on performance.

In 2015 we started to study the state of the art about SLA for security monitoring in clouds, as well as about evaluating security monitoring setups in clouds.

In 2015 we also studied the self-adaptation issues of security monitoring with two kinds of security monitoring components: a network intrusion detection system (NIDS), and a secured application-level firewall. Moreover a new approach to secure an application-level firewall has been proposed.

To experiment with both kinds of components, a prototype called SAIDS has been implemented in the OpenStack-based IaaS cloud testbed that was setup in 2014. The NIDS software used is Snort. The application-level firewall is based on Linux nftables and Open vSwitch. In order to study more complex security monitoring setups, SAIDS will be extended in 2016.

A preliminary evaluation of SAIDS has been published in the doctoral symposium of CCGrid 2015. A more complete evaluation of SAIDS as well as the evaluation of the application-level firewall will be done in 2016.

7.4. Greening Clouds

Participants: Maria Del Mar Callau Zori, Ismael Cuadrado Cordero, David Guyon, Sabbir Hasan Rochi, Yunbo Li, Christine Morin, Anne-Cécile Orgerie, Jean-Louis Pazat, Guillaume Pierre.

7.4.1. Energy-aware IaaS-PaaS co-design

Participants: Maria Del Mar Callau Zori, Anne-Cécile Orgerie, Guillaume Pierre.

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 conducted an extensive experimental evaluation of the effect of a range of Cloud infrastructure operations (start, stop, migrate VMs) on their computing throughput and energy consumption, and derived a model to help drive cloud reconfiguration operations according to performance/energy requirements. A publication on this topic is in preparation.

7.4.2. Energy-efficient cloud elasticity for data-driven applications

Participants: David Guyon, Anne-Cécile Orgerie, Christine Morin.

Distributed and parallel systems offer to users tremendous computing capacities. They rely on distributed computing resources linked by networks. They require algorithms and protocols to manage these resources in a transparent way for users. Recently, the maturity of virtualization techniques has allowed for the emergence of virtualized infrastructures (Clouds). These infrastructures provide resources to users dynamically, and adapted to their needs. By benefiting from economies of scale, Clouds can efficiently manage and offer virtually unlimited numbers of resources, reducing the costs for users.

However, the rapid growth for Cloud demands leads to a preoccupying and uncontrolled increase of their electric consumption. In this context, we will focus on data driven applications which require to process large amounts of data. These applications have elastic needs in terms of computing resources as their workload varies over time. While reducing energy consumption and improving performance are orthogonal goals, this internship aims at studying possible trade-offs for energy-efficient data processing without performance impact. As elasticity comes at a cost of reconfigurations, these trade-offs will consider the time and energy required by the infrastructure to dynamically adapt the resources to the application needs.

The master internship work of David Guyon on this topic has been presented at IEEE GreenCom 2015 [39]. This work will be continued during David's PhD thesis.

7.4.3. Energy-efficient and network-aware resource allocation in Cloud infrastructures

Participants: Ismael Cuadrado Cordero, Christine Morin, Anne-Cécile Orgerie.

Energy consumption is cloud computing has become a key environmental and economic concern. Our work aims at designing energy-efficient resource allocation for Cloud infrastructures. The ever-growing appetite of new applications for network resources leads to an unprecedented electricity bill, and for these bandwidth-hungry applications, networks can become a significant bottleneck. New algorithms have to be designed integrating the data locality dimension to optimize computing resource allocation while taking into account the fluctuating limits of network resources. Towards this end, we proposed GRaNADA, a semi-decentralized Platform-as-a-service (PaaS) architecture for real-time multiple-users applications. Our architecture geographically distributes the computation among the clients of the cloud, moving the computation away from the datacenter to save energy - by shutting down or downgrading non utilized resources such as routers and switches, servers, etc. - and provides lower latencies for users. GRaNADA implements the concept of micro-cloud, a fully autonomous energy-efficient subnetwork of clients of the same service, designed to keep the greenest path between its nodes. Along with GRaNADA, we proposed DEEPACC, a cloud-aware routing protocol which distributes the connection between the nodes. Our system GRaNADA targets services where the geographical distribution of clients working on the same data is limited - for example, a shared on-line document - or those services where, even if the geographical distribution of clients is high, the upload data communication to the cloud is small - for instance a light social network like Twitter. We compared our approach with two main existing solutions - replication of data in the edge and traditional centralized cloud computing. Our approach based on micro-clouds exhibits interesting properties in terms of QoS and especially latency. Simulations show that, using the proposed PaaS, one can save up to 75% of the spent network energy compared to traditional centralized cloud computing approaches. Our approach is also more energy-efficient than the most popular semi-decentralized solutions, like nano data centers. This work has been presented at IEEE GreenCom 2015 [18].

We also evaluated the suitability of using micro-clouds in the context of smart cities. We investigated the idea to build a local cloud on top of networking resources spread across a defined area and including the mobile devices of the users. This local cloud is managed by lightweight mechanisms in order to handle users who can appear/disappear and move. We used a scenario considering a platform for neighborhood services and showed that micro-clouds make better use of the network, reducing the amount of unnecessary data traveling through external networks. This work is currently under review for a conference.

7.4.4. Resource allocation in a Cloud partially powered by renewable energy sources

Participants: Yunbo Li, Anne-Cécile Orgerie.

We propose here to design a disruptive approach to Cloud resource management which takes advantage of renewable energy availability to perform opportunistic tasks. To begin with, the considered Cloud is mono-site (i.e. all resources are in the same physical location) and performs tasks (like web hosting or MapReduce tasks) running in virtual machines. This Cloud receives a fixed amount of power from the regular electric Grid. This power allows it to run usual tasks. In addition, this Cloud is also connected to renewable energy sources (such as windmills or solar cells) and when these sources produce electricity, the Cloud can use it to run more tasks.

The proposed resource management system needs to integrate a prediction model to be able to forecast these extra-power periods of time in order to schedule more work during these periods. Batteries will be used to guarantee that enough energy is available when switching on a new server working exclusively on renewable energy. Given a reliable prediction model, it is possible to design a scheduling algorithm that aims at optimizing resource utilization and energy usage, problem known to be NP-hard. The proposed heuristics will thus schedule tasks spatially (on the appropriate servers) and temporally (over time, with tasks that can be planed in the future).

This work is done in collaboration with Ascola team from LINA in Nantes. Two publications have been accepted this year on this topic for: SmartGreens 2015 [15] and IEEE GreenCom 2015 [21].

7.4.5. SLA driven Cloud Auto-scaling for optimizing energy footprint

Participants: Sabbir Hasan Rochi, Jean-Louis Pazat.

As a direct consequence of the increasing popularity of Internet and Cloud Computing services, data centers are amazingly growing and hence have to urgently face energy consumption issues. At the Infrastructure-as-a-Service (IaaS) layer, Cloud Computing allows to dynamically adjust the provision of physical resources according to Platform-as-a-Service (PaaS) needs while optimizing energy efficiency of the data center.

The management of elastic resources in Clouds according to fluctuating workloads in the Software-as-a-Service (SaaS) applications and different Quality-of-Service (QoS) end-user's expectations is a complex issue and cannot be done dynamically by a human intervention. We advocate the adoption of Autonomic Computing (AC) at each XaaS layer for responsiveness and autonomy in front of environment changes. At the SaaS layer, AC enables applications to react to a highly variable workload by dynamically adjusting the amount of resources in order to keep the QoS for the end users. Similarly, at the IaaS layer, AC enables the infrastructure to react to context changes by optimizing the allocation of resources and thereby reduce the costs related to energy consumption. However, problems may occur since those self-managed systems are related in some way (e.g. applications depend on services provided by a cloud infrastructure): decisions taken in isolation at given layer may interfere with other layers, leading whole system to undesired states.

We have defined a scheme for green energy management in the presence of explicit and implicit integration of renewable energy in datacenter [13]. More specifically we propose three contributions: i) we introduce the concept of virtualization of green energy to address the uncertainty of green energy availability, ii) we extend the Cloud Service Level Agreement (CSLA) language to support Green SLA introducing two new threshold parameters and iii) we introduce greenSLA algorithm which leverages the concept of virtualization of green energy to provide per interval specific Green SLA. Experiments were conducted with real workload profile from PlanetLab and server power model from SPECpower to demonstrate that, Green SLA can be successfully established and satisfied without incurring higher cost.

This work is done in collaboration with Ascola team from LINA in Nantes.

7.5. Energy-efficient Computing Infrastructures

Participants: Christine Morin, Anne-Cécile Orgerie, Martin Quinson.

7.5.1. Simulating the impact of DVFS within SimGrid

Participants: Christine Morin, Anne-Cécile Orgerie, Martin Quinson.

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 experimentation within the SimGrid simulation toolkit, by introducing a model of application energy consumption and enabling the use of Dynamic Voltage and Frequency Scaling (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. This work is done in collaboration with the Mescal team from LIG in Grenoble. A paper is currently under preparation on this work.

7.5.2. Simulating Energy Consumption of Wired Networks

Participants: Timothée Haudebourg, Anne-Cécile Orgerie.

Predicting the performance of applications, in terms of completion time and resource usage for instance, is critical to appropriately dimensioning 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 dimensioning 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 the 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.

In 2015, this simulator has been extended to integrate two more green levers: low power idle (IEEE 802.3az) and adaptive link rate. This work has been done during the internship of Timothée Haudebourg (L3 ENS Rennes) and a publication is currently under preparation.

7.5.3. Multicriteria scheduling for large-scale HPC environments

Participant: Anne-Cécile Orgerie.

Energy consumption is one of the main limiting factor for the design and deployment of large scale numerical infrastructures. The road towards "Sustainable Exascale" is a challenge with a target of 50 Gflops per watt. Energy efficiency must be taken into account and must be combined with other criteria like performance, resilience, Quality of Service.

As platforms become more and more heterogeneous (co-processors, GPUs, low power processors...), an efficient scheduling of applications and services at large scale remains a challenge. In this context, we will explore and propose a multicriteria scheduling model and framework for large scale HPC systems. Based on real energy measurements and calibrations, we will propose some performance and energy models and will build a multi criteria scheduler. Simulation on selected scenario will be explored and a prototype will be designed for ensuring experimental validation.

This work is done in collaboration with ROMA and Avalon teams from LIP in Lyon.

7.6. Decentralized and Adaptive workflows

Participants: Jean-Louis Pazat, Javier Rojas Balderrama, Matthieu Simonin, Cédric Tedeschi, Palakiyem Wallah.

7.6.1. Adaptive Workflows with Chemical Computing

Participants: Javier Rojas Balderrama, Matthieu Simonin, Cédric Tedeschi.

We have designed a high-level programming model based on the HOCL rule-based language to express workflow adaptation. It was specifically designed to support changes in the workflow logic at run time. This mechanism was implemented within the GinFlow software and experimented over the Grid'5000 platform. An article was just accepted for publication at the IPDPS 2016 conference.

7.6.2. Best-effort decentralized workflow execution

Participants: Jean-Louis Pazat, Cédric Tedeschi, Palakiyem Wallah.

We are currently proposing a simple workflow model for workflow execution in platforms with limited computing resources and services. The key idea is to devise a best-effort workflow engine that does not require a strong centralized orchestrator. Such a workflow engine relies on point-to-point cooperation between nodes supporting the execution. A minimalistic demonstrator of these concepts has been devised and implemented. Early experiments have been conducted on a single machine.

7.7. Experimental Platforms

Participants: Julien Lefeuvre, David Margery.

7.7.1. Contribution to Fed4FIRE testbed

Participants: Julien Lefeuvre, David Margery.

In Fed4FIRE, two key technologies have been adopted as common protocols to enable experimenters to interact with testbeds: Slice Federation Architecture (SFA), to provision resources, and Control and Management Framework for Networking Testbeds (OMF) to control them. In 2015, the main area of work has been the implementation of an SFA API to BonFIRE, still on-going. In the process, we wrote the reference documentation to write a new delegate for geni-tools, the reference implementation of SFA maintained by the GENI project office. This codebase has now been made public on github, in part because of our interactions with the code and suggested changes to ease writing new delegates. We have also contributed to the design of a service layer proxy mechanisms so that testbeds with http based APIs can be queried by any Fed4FIRE user using a standard authentications mechanism. The BonFIRE API has been made available through that mechanism, based on XML documents signed using the XML Signature specification.

TACOMA Team

6. New Results

6.1. Self-describing objects and tangible data structures

Participants: Nebil Ben Mabrouk, Paul Couderc [contact].

A development in the line of the composite objects (see section 3.3) 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, 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 analysing 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.

In the context of RFID system, the resiliency property of such data structures enables new information architecture and autonomous (offline) operation, which is very important for some RFID applications. We previously applied the self-describing objects approach to the waste management domain [1], which has shown to be a specially challenging situation for RFID. This challenge is found more generally in pervasive computing scenarios involving RFID reading in uncontrolled environments (see section 4.3).

Pervasive support for RFIDs.

We propose to apply our approach to improve the robustness of RFID inventories / batch checking: when many objects are read at once by an RFID reader, miss read are common and raise reliability and operational issues for applications. An innovative solution to this problem is to take advantage of the multiplicity of tags by leveraging them as a distributed memory shared by a logical group. In this way, it is possible to support error detection as well as information recovery. We proposed a flexible protocol to support robust EPC retrieval in adverse reading conditions. The proposed protocol uses erasure correcting techniques to enable error-free recovery of misread EPCs [2]. It is further customizable with respect to the rate of misread tags and application requirements. This work was the object of an Inria patent [0]. Fine-tuning the protocol parameters is still the object of on going experimentation in the context of the Pervasive_RFID project (see section 7.1.1).

[0] Patent filed in April 2015 - Inria 179

At the software level, RFID inventory reliability issue is usually addressed by anti-collisions mechanisms and redundancy mechanisms. Anti-collisions protocols limit the risk of data corruption when multiples tags have to reply to an inventory request. Redundancy is often implemented in RFID readers by aggregating the results of multiple inventory requests over a time frame, to give the tags multiple opportunities to reply. While useful, these strategies cannot ensure that a given inventory is valid or not (in other words, one or more tags may be missing without being noticed). In situations where we have to read large collection of objects of various types, the performance is difficult to predict but may still be adequate for a given application. For example, some application can tolerate missing some tags, provided that miss read probability could be characterized. In some cases, read reliability could be improved using mechanical approaches, such as introducing movements in objects or antenna to introduce *radio diversity* during read. Finally, distributed data structure can be used over a set of tags to be used to mitigate the impact of misread (by using data redundancy) and to help the reading protocol by integrating hints about the tag set collection being read.

We studied extensively by experimentation the behaviour of existing RFID solutions in the context of uncontrolled environment (meaning, random placement of tags on objects mixing various materials) in order to characterize their real-world performance regarding the parameters of such as tags numbers, density, frequencies, reader antenna design, dynamicity of objects (movements), etc. From these experimentations, we would like to identify the conditions that are favorable to acceptable performance, and the way where there are hopes of improvement with specific design for these difficult environments. These results should also allow improving the performance: high level integrity checks can guide low level operations by determining whether inventories are complete or not. This cross layer strategy enables faster and more efficient inventory protocols.

6.2. Interactions between connected objects in a Smart Building

Participants: Adrien Capaine, Yoann Maurel, Frédéric Weis [contact].

Tacoma group is focussed on the conception and implementation of innovative services for the Smart Home/Building. The range of considered services is broad: from "optimizing the energy consumption" to "helping users to find their way in a building". One of our goals is to build a pervasive platform with constrained performance and cost [7], without disrupting existing spaces. Within this idea, we explored in 2015 the services provided by different modes of interaction in a physical space between neighboring objects, and also between an object and a nearby user.

More precisely, we conducted some experiments with LEDs. Instrumented via a short distance radio interface, a lighting device becomes an unobtrusive connected object that is easy to integrate to a mesh network. A relevant aspect of this platform is the consideration of potential conflicts in data access offered by the connected objects. One of the first scenarios we considered is to operate an LED-based light path to guide the evacuation of a building in the case of a fire alarm. When our objective is to multiply the uses of LED devices ("go beyond lighting", see section 7.1.2), the question is then the priority of access to resources offered by the platform distributed in the environment. Specifically, we addressed the following issues (similar to some of the issues presented in section 3.2):

- How to prioritize the lighting functions (classic) and occasional (but priority) uses of the LED to help in the care of a fire alarm?
- How do you prioritize access to the objects and/or resources that carry these items?

6.3. Context computing for Smart Home

Participants: Yoann Maurel, Frédéric Weis [contact].

To provide services for smart Homes, automation based on pre-set scenarios is ineffective: human behavior is hardly predictable and application should be able to adapt their behavior at runtime depending on the context. We focused on recognizing user's activities to adapt applications behaviours. Our aim is to compute small pieces of context we called *context attributes*. Those context attributes are diverse, for example a presence in a room, the number of people in a room etc.

Building efficient and accurate context information using inexpensive and non-invasive sensors was and is still a great challenge 3.1 . We proved, through the use of dedicated algorithms and a layered architecture that it is achievable when the targeted Home is known - due to the specific and non automated calibration process we used. Among all the available theories, we used the Belief Function Theory (BFT) [8] [9] as it allows to express **uncertainty** and **imprecision**.

Context is computed by a chain a tasks as illustrated in figure 5 :

- The transition between a raw sensor value and a belief function is made through the use of a belief model which maps a sensor value to a belief function. A belief function represents the degree of belief associated to each possible value of the context attribute.
- Then a set of belief functions (corresponding to a set of sensors) can be combined (fused).
- Finally the system can decide what is the "best" value for the context attribute.

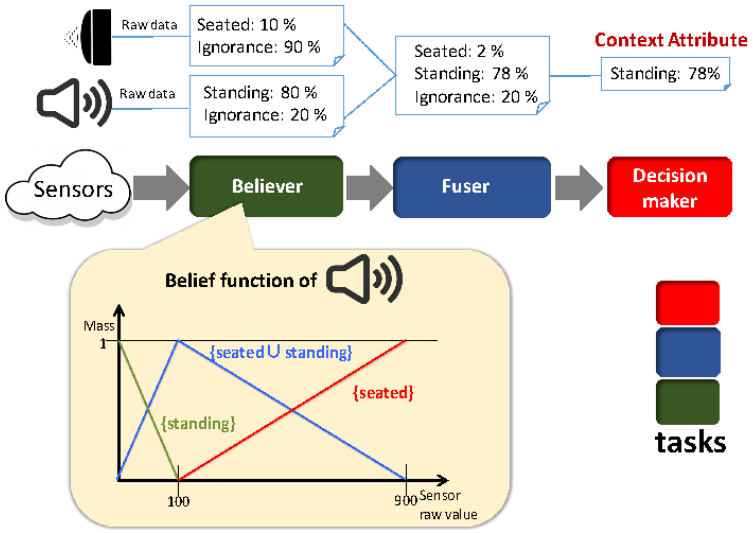


Figure 5. Context attribute computation

Alleviating the complexity of the platform configuration and maintenance is a prerequisite for the adoption of Smart-Home environments by consumers. Currently the BFT theories requires a huge calibration process. We focussed our efforts on the semi-automated building of belief functions, required by the theory, that have to be provided by each sensor.

Automated configuration of sensors.
The belief model is provided to the platform by us and a component is in charge of transforming a sensor value in a *belief function*. The fine tuning of a belief function can be a tedious task. It must be done by a specialist who understands the belief function theory and knows the behavior of the sensors. The model is often built iteratively by experimenting. This may take several hours or days. Moreover, this method is directly connected to the output of each sensor. Biased and noisy measures can cause major modifications on the resulting beliefs.

Ideally, the calibration of the model should be as automatic as possible (few interaction with the user during calibration). The person setting up the sensors should not have to understand the belief function theory. We proposed to generate our belief model from a training set of sensor data. We mainly focused on k-nearest neighbors (KNN) algorithm [6]. We used a training data set to compute the presence belief model. We acquired a set of data with someone present in the experimentation room and a second data set with nobody in the room, which gives us a labelled data set. This principle is illustrated in figure 6 .

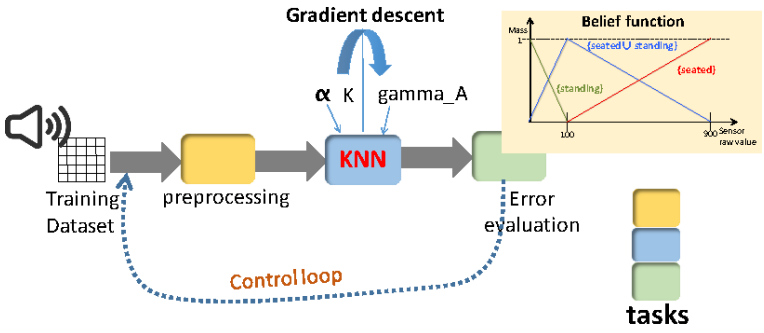


Figure 6. Sensor autocalibration

6.4. Design of a framework for distributed pervasive environments

Participants: Adrien Capaine, Yoann Maurel [contact], Frédéric Weis.

Pervasive environment brings into play complex interactions between a large number of heterogeneous entities: computing units executing third-party applications delivering multiple services to users, with various (sometimes conflicting) requirements, based on the information provided by dynamically (un)available smart object or sensors. The development of pervasive application is consequently hard and must be supported by architectures and frameworks that propose solution to manage the heterogeneity, to organize the interaction of distributed entities, to support the dynamic discovery of the entities, to ensure the privacy of collected data and inferred context, to organize and structure information sharing, and to enforce access control over data and entities.

To alleviate the development of such application (see section 3.4), we worked on a distributed pervasive environment made of several processing nodes (or gateway) managing interacting Smart Spaces (*i.e* a room, a corridor etc.). A Smart Space contains one or more nodes that coordinate to provide services to users. A node is a low cost computing unit with constrained performances. Each node is responsible for the management of entities and services available in their close proximity: they dynamically discover available devices and source of information, computes contextual information and offer services to nearby users. Nodes are organized hierarchically: in each space one *supervisor node* is responsible for coordination between nodes (*e.g.* managing conflicting requirements and enforcing global policies) and communication with neighboring Smart Spaces. The whole environment (a Smart Building or a Smart Home) is controlled by a master node that distributes policy between Smart Spaces to provide global services (*e.g.* global energy management). Data are stored and processed by nodes as close as possible to the users in order to enforce data privacy (see section 2.1).

Each node supports the execution of several services and application. To help the development of these services, applications are built on top of a framework (**Matriona**) running on each gateways.

Matriona proposes a unified representation of the concepts manipulated by applications in order to hide the heterogeneity of technologies to the application. We rely on the concept of *resource*. A Matriona's resource is quite similar to a REST resource but is not identical: HTTP protocol is only used for remote call; the structure of a resource is constrained; a resource can be dynamically discovered; it can provide notification (PUSH operation); it is uniquely identified in the environment; it is any object used or shared by applications, by bridges, or by the system itself: a device, a room, a platform, a user or a contextual information; it is implemented as a standard object and has a type specified by its interfaces (annotated Java interface). Types describe the data and the operations on resources. *CRUD* (**C**reate, **R**ead, **U**pdate, **D**elete) and *PUSH* operations are used to represent the semantic of operations. Specifying the semantic of operations allows to operate some

generic processing (*e.g.* gathering all the information provided by a resource) on resources without knowing their types.

Matriona enables dynamic discovery between nodes and inter-platform communication between the applications and resources. The platforms automatically discovers other platforms using a discovery mechanism. Each platform enables the discovery and the use of its resources to other platforms' applications. Remote resources are using exactly the same API as their internal counterparts: using remote resources is completely transparent to the application. The data is serialized / deserialized automatically by the platform during the call. Calling remote resources induces a performance cost equivalent to the use of a traditional REST resource.

Matriona allows to dynamically add new properties and new behaviors to a resource using decorators. Resources are built using multiple layers in the same way as "Russian dolls". Each layer is responsible for implementing a specific behavior such as retrieving, conversion operation or adding new properties (*e.g.* localization). For instance, the implementation of a thermometer resource will consist of 1) a core layer providing standard information (id, date of creation, the groups to which it belongs); 2) a protocol layer able to communicate with the real devices; 3) a conversion layer (Kelvin to Celsius). The application interacts with the top layer that exposes all or part of information and treatments offered by lower layers. While some layers are static (part of the resource declaration) and cannot be removed, most can be added afterward by the applications. It creates a virtual resource composed of the original resource and the new layers. This virtual resource can be used and discovered as any other resources.

Matriona allows to organize the information. Resource may reference other resources, for example the localization of a "thermometer resource" refers to the resource representing the room in which it is located. The value of the property is the *id* of the referenced resource. This allows applications to easily find resources and their interactions. It is also possible to create composite resources using aggregation mechanisms provided by the framework. The virtual resource can be used directly by applications as any other resources.

Matriona provides a basic language queries. Applications use resources directly or send queries to the platform registry. The query language allows to apply filters and to aggregate data on available resources. A request is represented by a specific URL. For instance, the mean temperature of thermometers in the whole meeting rooms of a building can be obtained using the URL `/*/*/$ location/meeting_room/temperature!mean`. The query language can be extended by providing new decorators and new filters.

Matriona provides access management: each resources belongs to one or many groups. The groups are defined when the resource is created or during its lifecycle by the owner of the resource. Groups gather applications that share the same permissions on access resources. Groups are managed by a "group owner" that can limit members permissions. Permissions describe the ability of an application to read, write, update, delete, manage or lock a resource. Resource locking avoids conflicting requests to be performed by different applications. Locks are given to applications for a fixed period of time. A resource can always be unlocked by the platform itself or by "critical" applications (*e.g.* emergency fire alarm, see section 6.2).

Matriona allows applications to extend resource properties and to share these meta-information with others. Each application can add new information to a given resource. Tagged information are available only for the application and its group. Meta-information are stored by the platform and associated to the resource until the latter is destroyed. This mechanism allows application to easily share information on the resource they used. For instance, this can be used to retrieve previously used resources or to rate the quality of service provided by a given resource. For the application, meta-information are part of the resource. It is then possible for an application to only use resources that have been approved by other applications of their group. This mechanism can also be used by application to add some task-relevant information (*e.g.* a medical application can tag resources that have been used by a patient).

6.5. Towards Metamorphic Housing: the on-demand room

Participant: Michele Dominici [contact].

This research activity is supported by Fondation Rennes 1 through the chair "Smart Home and Innovation", since January 2014. This activity is centered on the concept of metamorphic housing (see section 4.2). During the first year, we had identified the goals of the research project, also taking into account the trends of future housing industry, provided by the enterprises and public authorities that support the chair. We also had identified a case study, the on-demand room, to be displayed as the main application of the research results in scientific communications and vulgarization. It consists in a space that is physically shared by a small group of apartments, but is assigned for the sole use of one or few particular ones at the time. The room is designed so as to make occupants feel they did not leave their apartment at all. They seamlessly move from their dwelling to the on-demand room and conversely, without noticing the difference, as the room adapts to their preferences. During 2015, second year of the chair, we organized our work following two main axes: *(i)* solving the research problems, illustrated in the rest of this section; *(ii)* demonstrating the results using mixed reality as combination of virtual reality and off-the-shelf domotic devices, described in section 5.2.2 .

The research problems underlying the on-demand room are numerous: we illustrated them in the research report "A Case Study of Metamorphic Housing: The on-Demand Room" [3]. We started by addressing the problems associated with the goal of "plugging" the room into different apartments. This requires to dynamically change the rights to control and customize the room's equipment, including lights, appliances, heating, ventilation and air conditioning systems (HVAC), etc. This must be done in a transparent fashion, so that off-the-shelf devices and appliances can be used.

To solve these problems, we started a collaboration with the DIVERSE team [0]. The goal is to use the Kevoree [0] software framework to dynamically reconfigure the networks and domotic system of the room and of the apartments. When the on-demand room is owned by an apartment, their computer networks are interconnected; appliances, sensors and controllers in the room and the apartment can communicate with each others; devices reflect user preferences. Kevoree will enable these reconfigurations by running on key appliances and dynamically adapting and customizing their behavior to the owner of the on-demand room.

As part of the collaboration, some research goals have already been identified. The underlying challenges will be addressed and the results will be integrated in a comprehensive mixed reality demonstrator. This will represent the final iteration of the ongoing demonstration process, illustrated in the platform section (for more details, see 5.2.2).

[0] http://diverse.irisa.fr/
[0] http://kevoree.org/

DREAM Project-Team

7. New Results

7.1. Simulator-based decision support

Participants: Philippe Besnard, Marie-Odile Cordier, Anne-Isabelle Graux, Christine Largouët, Véronique Masson, Laurence Rozé.

7.1.1. Ecosystem model-checking for decision-aid

Former studies of ecosystem modelling have concentrated on temporal modelling. In recent studies we have focussed on the formalization of spatial diffusion of a prey-predator trophic network composed of weeds and ground beetle. For this purpose, an approach coupling landscape representation and population models has been used. A reaction-diffusion model was developped through the synchronization ability of timed-automata. The agronomical rules of beetle migration and weeds diffusion have been translated into communications between timed automata. Landscapes have been simulated and can be evaluated thanks to landscape-metrics distance. The optimization aims to maximize the ground beetle abundance while minimizing the use of pesticides. The model obtained in this first study is quite complex but preliminary results are beeing studied.

7.1.2. Controler synthesis for optimal strategy search

Similarly to previous work, this approach relies on a qualitative model of a dynamical system. The problem consists in finding a strategy in order to help the user achieveing a specific goal. The model is now considered as a timed game automata expressing controllable and uncontrollable actions. The strategy represents the sequence of actions that can be performed by a user to reach a particular state (in case of a reachability problem for instance). A first approach based on a "generate and test" method has been developped for the marine ecosystem example [86].

Recently, we generalized the work of Yulong Zhao applied in the context of a dairy production system [87] to the planning domain. The planning task consists in selecting and organizing actions in order to reach a goal state in a limited time and in an optimal manner, assuming actions have a cost. We propose to reformulate the planning problem in terms of model-checking and controller synthesis on interacting agents such that the state to reach is expressed using temporal logic. We have chosen to represent each agent using the formalism of Priced Timed Game Automata (PTGA). PTGA is an extension of Timed Automata that allows the representation of cost on actions and uncontrollable actions. Relying on this domain description, we define a planning algorithm that computes the best strategy to achieve the goal. This algorithm is based on recognized model-checking and synthesis tools from the UPPAAL suite. The expressivity of this approach is evaluated on the classical *Transport Domain* which is extended in order to include timing constraints, cost values and uncontrollable actions. This work has been implemented and performances evaluated on benchmarks.

7.1.3. A datawarehouse for simulation data

In previous work we have proposed a datawarehouse architecture to store the huge data produced by deep agricultural simulation models [50]. This year, we have worked on hierarchical skyline queries to introduce skyline queries in a datawarehouse framework. 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 [52]. It provides the user with navigation operators along the dimensions hierarchies (i.e. specialize / generalize) while ensuring an online calculation of the associated skyline.

Anne-Isabelle Graux, on leave from INRA (National Institute for Agronomical Research), is working on an adaptation and extension of this method for storing the simulation results of a comprehensive farm model named MELODIE [53]. The new datawarehouse will enable the analysis of simulation results within dynamic preferences, related to grassland management for instance, for identifying the data satisfying the best compromises with respect to possibly inconsistent criteria.

7.1.4. Post-mining classification rules

We consider sets of classification rules with quantitative and qualitative attributes inferred by supervised machine learning, as in the framework of the Sacadeau project. Our aim is to improve the human understanding of such sets of rules. First, we consider quantitative attributes in rules that often contain too many intervals which are difficult to intepret. We propose two algorithms to merge some of these intervals in order to get more understandable rules. These algorithms take into account the final rule quality. We are also working on formalizing what could be the quality of a set of rules. There are lots of studies about the quality of one rule but very few about the quality of the whole set of rules and this is still an issue.

7.2. Data Mining

Participants: Marie-Odile Cordier, Yann Dauxais, Serge Vladimir Emteu Tchagou, Clément Gautrais, Thomas Guyet, Yves Moinard, Benjamin Negrevergne, René Quiniou, Laurence Rozé, Alexandre Termier.

7.2.1. Sequential pattern mining with intervals

In previous work, we developed a framework for sequential pattern mining with intervals [3]. It has been applied in various application (care-pathways, customer relationship management databases [35], etc.).

This year we explored chronicle mining algorithms for mining care-pathways (see section 9.1.1 , for an applicative context). Chronicles are alternative patterns for representing temporal behaviors [58]. A chronicle can be briefly defined as a set of events linked by constraints indicating the minimum and maximum time elapsed between two events. A care-pathway contains point-based events (e.g. surgery) and interval-based events (e.g. drug exposures). A chronicle can express such a complex temporal behaviour, for instance: *The patient was exposed to a drug X between 1 and 2 years, he met his doctor between 400 to 600 days after the beginning of the exposure and, finally, he was hospitalized.*

The first algorithm we worked on [23] is an adaptation of existing chronicle mining algorithms [55], [63] to mine the complete set of frequent chronicles from a collection of care-pathways. This algorithm uses the search-space browsing strategy of HDCA [55] and the support evaluation of CCP-Miner [63]. As the complete set of chronicle is huge, we also proposed an incomplete algorithms based on the original simplifications of [58]. These algorithms were implemented and evaluated on real and simulated datasets.

We also investigated discriminant chronicles mining which consists in extracting the chronicles that are α times more frequent in a database $\mathcal{D}_+$ than in a database $\mathcal{D}_-$. Mining discriminant chronicles is very useful to discover the features of care-pathways that are related, for instance, to a specific disease. Our approch has been implemented and is under evaluation.

7.2.2. Multiscale segmentation of satellite image time series

Satellite images enable the acquisition of large-scale ground vegetation information. Images have been recorded for several years with a high acquisition frequency (one image every two weeks). Such data are called satellite image time series (SITS). Several articles were published this year and they correspond to past work on algorithms and method to analyse SITS.

In [11], 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 also explored the supervised classification of SITS using classification trees for time-series [27] by implementing a parallelized version of this algorithm. Next, we explored the adaptation of the object-oriented segmentation to SITS. The object-oriented segmentation is able to segment images based on segment uniformity. We proposed a measure for time-series uniformity and applied the adapted algorithm on large multivariate SITS of Senegal [10].

Third, we presented an supervised approach to extract features from classified satellite images to analyse urban sprawl [28]. In this work, we have satellite images at only two dates, and the objective is to identify characteristics that can foster or prevent changes.

Our satellite images analysis approaches are used in two applicative contexts: understanding urban sprawl and analyzing drought in Senegal. Analysis of urban sprawl was a collaborative work with collegues in remote sensing, in landscapes analysis and in economical modelling. Our collective contribution was published in a book of the PDD2 [0] program [38]. Analysis of drought in Senegal is a long term collaboration with H. Nicolas (INRA/SAS) that we would like to continue in a collaboration with A. Fall (Université of Dakar) to confront our results with ground observations.

7.2.3. Analysis and simulation of landscape based on spatial patterns

Researchers in agro-environment need a great variety of landscapes to test their scientific hypotheses using agro-ecological models. Real landscapes are difficult to acquire and do not enable the agronomists to test all their hypotheses. 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 landscapes that reproduce realistic interface properties between parcels. This approach consists of the extraction of spatial patterns from a real geographic area and the use of these patterns to generate new "realistic" landscapes. It 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.), in a specific geographic area.

In past years, we worked on the exploration of graph mining techniques, such as gSPAN [85], to discover the relevant spatial patterns present in a spatial-graph. We assume that the set of the frequent graph patterns are the characterisation of the landscape. Our remaining challenge was to simulate new realistic landscapes that reproduce the same patterns.

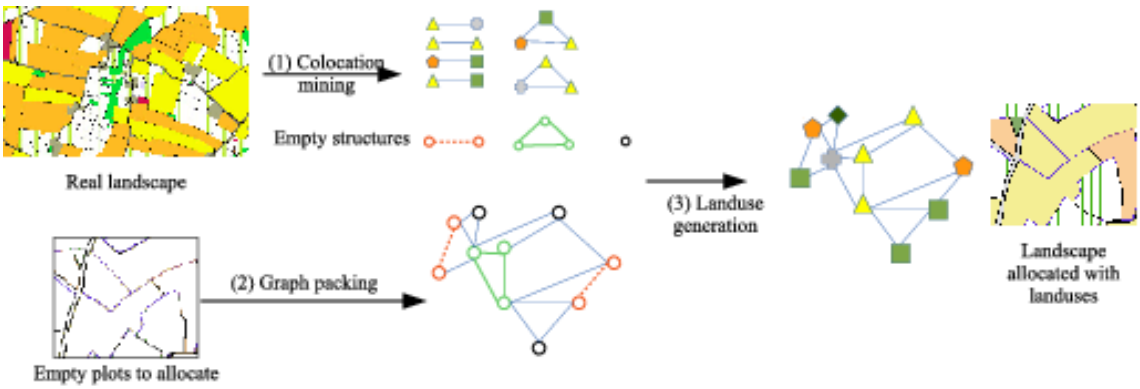


Figure 1. Simulation process in three steps: 1) characteristic graph-patterns mining, 2) graph packing of the cadastral landscape and 3) crop allocation.

[0]PDD2: Paysage Developpement Durable/Landscape Sustainable Developpent

We have formalized the simulation process as a graph packing problem [66]. The process is illustrated by Figure 1 . Solving instances of the general graph packing problem has a high combinatorics and no efficient algorithm can solve it. We proposed an ASP program to tackle the combinatorics of the graph packing and to assign the land use considering some expert knowledge. Our approach combines the efficiency of ASP to solve the packing issue and the simplicity of the declarative programming to take into account expert contraints on the land use. Contraints about the minimum surface of crops or about the impossibility of some crops colocation can be easily defined. This work have been presented at the conference RFIA and an extended version has been published in the Revue d'Intelligence Artificielle (RIA) [13].

In addition, we are collaborating with J. Nicolas (EPI Dyliss) to improve the efficiency of our first programs. The improvements are based on symmetry breaking of ASP programs. To this end, we proposed a simplified encoding of the graph patterns using spanning trees and used automorphism detection in graph patterns to automatically encodes symmetry breakings. Intensive evaluation of our encoding shown that this improvement enable to tackle significantly larger graphs than early programs did. This work will be soon submitted to a high ranking conference.

7.2.4. Mining with ASP

In pattern mining, a pattern is considered interesting if it occurs frequently in the data, i.e. the number of its occurrences is greater than a fixed given threshold. As non informed mining methods tend to generate massive results, there is more and more interest in pattern mining algorithms able to mine data considering some expert knowledge. Though a generic pattern mining tool that could be tailored to the specific task of a data-scientist is still a holy grail for pattern mining software designers, some recent attempts have proposed generic pattern mining tools [61] for itemset mining tasks. In collaboration with Torsten Schaub, we explore the ability of a declarative language, such as Answer Set Programming (ASP), to solve pattern mining tasks efficiently. In 2011, Jarvisälo proposed a first attempt devoted to itemset mining [64]. In Dream, we are working on sequential pattern mining, which is known to be more challenging than itemset mining and which has been also recently considered by constraint programming approaches [76].

We have worked on encoding in ASP most of sequential pattern mining tasks: sequences with constraints (gaps, maximum length, etc.), closed/maximal patterns, emergent sequences. Our first result is to show that ASP is suitable for encoding such complex pattern mining tasks. The experimental results show that our purely declarative approach is less efficient than constraint programming approaches [36]. Nonetheless, it is suitable to be blend with intensive knowledge. The challenge is now to show that our ASP framework can extract the meaningful patterns that other approaches loose in the overwhelming amount of sequential patterns.

A first attempt has been done in this direction in collaboration with J. Romero from the University of Potsdam. We used the system ASPRIN to define preferences on patterns. Defining preferences on patterns is also a classical approach to select the most interesting patterns. Some classical preferences on sequential patterns have been defined and the ASPRIN system is used to extract the preferred patterns according to one preference or a combination of preferences (skypatterns [81])

This work will be soon submitted to a high ranking international conference.

7.2.5. Mining time series

Monitoring cattle. Following the lines of a previous work [79], we are working on a method for detecting Bovine Respiratory Diseases (BRD) from behavioral (walking, lying, feeding and drinking activity) and physiological (rumen temperature) data recorded on feedlot cattle being fattened up in big farms in Alberta (Canada). This year, we have especially worked on multivariate sensor data analysis, especially on the evaluation of different combinations of sensors for determining the best configuration and parameter setting. This work was part of Afra Verena Mang's master thesis defended in september 2015 [73]. Two papers are in preparation.

SIFT-based time-series symbolisation Time series classification is an application of particular interest with the increase of such data. Computing the distance between time-series is time consuming. An abstract representation of time-series that accurately approximates distances between time-series and makes easier

their comparison is highly expected. In [17], we proposed a time series classification scheme grounded on the SIFT framework [70] adapted to time series. The SIFTs feed a Bag-of-Words representation of time-series. We have shown that this framework efficiently and accurately classifies time series, despite the fact that BoW representation ignores temporal order.

Mining sequential patterns from multimedia data Analyzing multimedia data to extract knowledge is a challenging problem due to the quantity and complexity of such data. Finding recurrent patterns is one method to structure and segment the data. In a collaboration with the EPI LinkMedia, we have proposed audio data symbolization and sequential pattern mining methods to extract patterns from audio streams. Experiments show this the task is hard and that the symbolization is a critical step for extracting relevant audio patterns [29].

7.2.6. Mining customer data for predicting and explaining attrition

Predicting customer defection in a retail context is difficult because, in most situations, the customer does not leave the store totally (there is no contract break as with banks or phone operators). We have proposed a new pattern model for representing the evolution of an individual customer purchase behavior that enables to early detect and to explain customer attrition. In particular, this model enables the analyst to determine which important kinds of product receives less and less attention from the customer. Thus, this model provides actionable knowledge at an individual scale that lets the retailer trigger targeted marketing actions to counter attrition. A poster has been submitted to the EBDT conference. This work has been performed during Clément Gautrais's master [59] and will be further investigated and extended during his PhD.

7.2.7. Mining energy consumption data

Machine tools in companies consume a lot of energy (before, during and after producing worked pieces). This year, we are beginning to work, with the start-up Energiency, on mining machine tool energy consumption data in order to propose energy savings to the companies. Firstly, we try to determine, according to the analyzed company, which data-mining algorithm should be used and which is the best configuration and parameter setting. Then, we aim to extract actions rules from patterns to help companies to consume less energy.

7.2.8. Trace reduction

One problem of execution trace of applications on embedded systems is that they can grow very large, typically several Gigabytes for 5 minutes of audio/video playback. Some endurance tests require continuous playback for 96 hours, which would lead to hundreds of Gigabytes of traces, that current techniques cannot analyze. We have proposed TraceSquiz, an online approach to monitor the trace output during endurance test, in order to record only suspicious portions of the trace and discard regular ones. This approach is based on anomaly detection techniques. Our detailed experiments have shown that our approach has a good anomaly detection performance, and can reduce the size of an output trace by an order of magnitude [24]. Serge Emteu successfully defended his PhD about this work on the 15/12/2015 [5].

7.3. Causal reasoning and argumentation

Participants: Philippe Besnard, Louis Bonneau de Beaufort, Marie-Odile Cordier, Yves Moinard.

7.3.1. Searching for explanations from causal relations and ontology for argumentation

We have continued our work on reasoning (precisely search for explanations) from causal relations and ontology [48]. We resort to a well-known model [49] in computational argumentation in order to provide some structure to the collection of potential explanations given by our causal formalism. We have developed a case study, namely the Xynthia storm case, (February 2010, western France, trial September 2014) for which there exists a huge amount of data from various official reports. We have implemented an ASP program which thereby provides another application, besides those already mentioned: mining and landscape simulation, for ASP.

7.3.2. Cognitive maps and Bayesian causal maps

Cognitive map is a qualitative decision model which is frequently used in social science and decision making applications. This model allows to easily organize individuals' judgments, thinking or beliefs about a given problem in a graphical representation containing different concepts and influences between them. However, cognitive maps cannot model uncertainty within the variables and provides only deductive reasoning (predicting an effect given a cause). In [37], we show how to translate the knowledge represented in cognitive maps in the form of arguments and attack relations among them. Given a decision problem, we propose to build, first, a cognitive map by eliciting knowledge from experts and then to transform it into a weighted argumentation framework (WAF for short) for ensuring efficient reasoning. Another contribution concerns enriching the WAF obtained from a given cognitive map for dealing with dynamics through the consideration of a varying set of observations.

Cognitive maps and Bayesian networks are useful formalisms to address knowledge representation. Cognitive maps are powerful graphical models for gathering or displaying knowledge but while offering an easy means to express individuals judgments, drawing inferences remains a difficult task. Bayesian networks are widely used for decision making processes that face uncertain information or diagnosis but are difficult to elicitate. To take advantage of both formalisms and to overcome their drawbacks, Bayesian causal maps (BCM) were developed [75]. In [6], we propose to start from a causal map to construct the model and then set the conditional probabilities. Once the common causal map (CM) is built we can transform it into a BCM which combines causal modeling techniques and bayesian probability theory. We have developed a complete framework and applied it on a real problem in an environmental context. The implemented decision facilitating tool enables the representation of different shellfish dredgers views about their activity as well as the test of different fishery management scenarios.

HYBRID Project-Team

7. New Results

7.1. 3D User Interfaces

7.1.1. Novel 3D Interactive Techniques

THING: Introducing a Tablet-based Interaction Technique for Controlling 3D Hand Models Merwan Achibet, Anatole Lécuyer and Maud Marchal

The hands of virtual characters are highly complex 3D models that can be tedious and time-consuming to animate with current methods. We introduced the *THING* [17], a novel tablet-based approach that leverages multi-touch interaction for a quick and precise control of a 3D hand's pose 2 . The flexion/extension and abduction/adduction of the virtual fingers can be controlled for each finger individually or for several fingers in parallel through sliding motions on the surface of the tablet. We designed two variants of THING: **(1)***MobileTHING*, which maps the spatial location and orientation of the tablet to that of the virtual hand, and **(2)***DesktopTHING*, which combines multi-touch controls of fingers with traditional mouse controls for the global position and orientation of the hand model. We compared the usability of THING against mouse-only controls and a data glove in two controlled experiments. Results show that DesktopTHING was significantly preferred by users while providing performance similar to data gloves. Together, these results could pave the way to the introduction of novel hybrid user interfaces based on tablets and computer mice in future animation pipelines. This work was done in collaboration with Géry Casiez (Inria team MJOLNIR).

Figure 2. THING enables the control of 3D hand models (in blue) by sliding fingers along sliders arranged in a morphologically-consistent pattern on the tablet's screen. This creates a strong correspondence between user's input and pose of the controlled hand. Here, the user closes the virtual hand and then points the index finger.

Plasticity for 3D User Interfaces: New Models for Devices and Interaction Techniques Jérémy Lachoche and Bruno Arnaldi

We have introduced new models for device and interaction techniques to overcome plasticity limitations in Virtual Reality (VR) and Augmented Reality (AR) [26]. We aimed to provide developers with solutions to use and create interaction techniques that fit to the 3D application tasks and to the input and output devices available. The device model describes input and output devices and includes capabilities, limitations and representations in the real world. We also propose a new way to develop interaction techniques with an approach based on PAC and ARCH models [43]. These techniques are implemented independently from the specific devices used thanks to the proposed device model. Moreover, our approach aims to facilitate the portability of interaction techniques over different target OS and 3D frameworks. This work was done in collaboration with Thierry Duval (Lab-STICC), Éric Maisel (ENIB) and Jérome Royan (IRT B-Com).

Dealing with Frame Cancellation for Stereoscopic Displays in 3D User Interfaces Jérémy Lacoche, Morgan Le Chénéchal, Valérie Gouranton and Bruno Arnaldi

We explored new methods to reduce ocular discomfort when interacting with stereoscopic content, focusing on frame cancellation [27]. Frame cancellation appears when a virtual object in negative parallax (front of the screen) is clipped by the screen edges; stereopsis cue lets observers perceive the object popping-out from the screen while occlusion cue provides observers with an opposite signal. Such a situation is not possible in the real world. This explains some visual discomfort for observers and leads to a poor depth perception of the virtual scene. This issue is directly linked to the physical limitations of the display size that may not cover the entire field of view of the observer. To deal with these physical constraints we introduce two new methods in the context of interactive applications. The first method consists in two new rendering effects based on progressive transparency that aim to preserve the popping-out effect of the stereo. The second method focuses on adapting the interaction of the user, not allowing him to place virtual objects in an area subject to frame cancellation. This work was done in collaboration with Sébastien Chalmé (IRT B-Com), Thierry Duval (Lab-STICC) and Éric Maisel (ENIB).

7.1.2. Understanding Human Perception in VR

Distance Estimation in Large Immersive Projection Systems, Revisited Ferran Argelaguet and Anatole Lécuyer

When walking within an immersive projection environment, accommodation distance, parallax and angular resolution vary according to the distance between the user and the projection walls which can influence spatial perception. As CAVE-like virtual environments get bigger, accurate spatial perception within the projection setup becomes increasingly important for application domains that require the user to be able to naturally explore a virtual environment by moving through the physical interaction space. In this work we performed two experiments which analyze how distance estimation is biased when accommodation distance, parallax and angular resolution vary [23]. The experiments were conducted in a large immersive projection setup with up to ten meter interaction range. The results showed that both accommodation distance and parallax have a strong asymmetric effect on distance judgments. We found an increased distance underestimation for positive parallax conditions as the accommodation-convergence difference increased. In contrast, we found less distance overestimation for negative and zero parallax conditions. Our findings also showed that angular resolution has a negligible effect on distance estimation. This work was done in collaboration with Anne-Hélène Olivier (MIMETIC) and Gerd Bruder (University of Hamburg).

Virtual Proxemics: Locomotion in the Presence of Obstacles in Large Immersive Projection Environments Ferran Argelaguet, Anatole Lécuyer

In the real world we navigate with ease by walking in the presence of obstacles, we develop avoidance strategies and behaviors which govern the way we locomote in the proximity of physical objects and other persons during everyday tasks. With the advances of virtual reality technology, it becomes important to gain an understanding of how these behaviors are affected in a virtual reality application. In this work, we analyzed the walking and collision avoidance behavior when avoiding real and virtual static obstacles [19]. In order to generalize our study, we considered both anthropomorphic and inanimate objects, each having his virtual and real counterpart. The results showed that users exhibit different locomotion behaviors in the presence of real and virtual obstacles, and in the presence of anthropomorphic and inanimate objects. Precisely, the results showed a decrease of walking speed as well as an increase of the clearance distance (i. e., the minimal distance between the walker and the obstacle) when facing virtual obstacles compared to real ones. Moreover, our results suggest that users act differently due to their perception of the obstacle: users keep more distance when the obstacle is anthropomorphic compared to an inanimate object and when the orientation of anthropomorphic obstacle is from the profile compared to a front position. We discussed implications on future large shared immersive projection spaces. This work was done in collaboration with Anne-Hélène Olivier (MIMETIC), Julien Pettré (MIMETIC) and Gerd Bruder (University of Hamburg).

7.1.3. Sports and Virtual Reality

A Methodology for Introducing Competitive Anxiety and Pressure in VR Sports Training Ferran Argelaguet and Anatole Lécuyer

Athletes' performance is influenced by internal and external factors, including their psychological state and environmental factors, especially during competition. As a consequence, current training programs include stress management. In this work, we explored whether highly immersive systems can be used for such training programs [11]. First, we proposed methodological guidelines to design sport training scenarios both on considering the elements that a training routine must have, and how external factors might influence the participant. The proposed guidelines are based on flow and social-evaluative threat theories. Second, to illustrate and validate our methodology, we designed an experiment reproducing a 10m Olympic pistol shooting competition 3 . We analyzed whether changes in the environment are able to induce changes in user performance, physiological responses and the subjective perception of the task. The simulation included stressors in order to raise a social-evaluative threat, such as aggressive public behavior or unforced errors, increasing the pressure while performing the task. The results showed significant differences in the user behavior and in their subjective impressions, trends in the physiological data were also observed. Taken together our results suggest that highly immersive systems could be further used for training systems in sports. This work was done in collaboration with Frank Multon (MIMETIC).

Figure 3. The proposed methodology was illustrated and evaluated in a virtual Olympic shooting experiment. The experiment was conducted in a wide immersive projection system being able to enclose a ten meter wide shooting range with six virtual opponents and one participant.

7.1.4. Experiencing the Past in Virtual Reality

An Immersive Virtual Sailing on the 18 th -Century Ship Le Boullongne Jean-Baptiste Barreau, Florian Nouviale and Valérie Gouranton

This work is the result of the collaboration between historians and computer scientists whose goal was the digital reconstitution of "Le Boullongne", an 18th-century merchant ship of "La Compagnie des Indes orientale" [12]. This ship has now disappeared and its reconstitution aims at understanding on-board living conditions. Three distinct research laboratories have participated in this project so far. The first, a department of naval history, worked on historical documents, especially the logbooks describing all traveling events of the ship. The second, a research laboratory in archeology, archaeoscience and history, proposed a 3D model of the ship based on the original naval architectural plans. The third, a computer science research laboratory, implemented a simulation of the ship sailing in virtual reality. This work focuses on the reconstitution of the ship in virtual reality, aiming at restoring a realistic interactive naval simulation: the 3D model of the ship has been integrated in an ocean simulation, with a physical rendering of the buoyancy. The simulation allows a user to walk around on the ship, at a scale of 1:1, and even steer it through a natural interaction. Several

characteristics of the simulation reinforce the sensation of being on-board: (1) A sonic environment mixing spatialized sounds (gulls flying, a whale swimming, wood cracking, cannons firing) and global soundscape (ocean and wind). (2) The meteorology of the simulation is dynamically modifiable; the user can increase the swell height and speed. The global illumination and wind sound vary in accordance with these parameters. The buoyancy simulation entails realistic movements of the ship. (3) Several interactions are proposed allowing the user to steer the ship with his/her hand, walk around on the ship, fire the cannons, and modify the weather. (4) Three animated sailors accompany the user in his/her sailing experience. They are wearing realistic period costumes. The immersive simulation has allowed historians to embark on "Le Boullongne" and to better understand how life was organized on-board. It has also been presented at several public exhibitions, in CAVE-like structures and HMD. This work was done in collaboration with Ronan Gaugne (Univ. Rennes 1), Yann Bernard (CReAAH) and Sylviane Llinares (CERHIO, UBS Lorient).

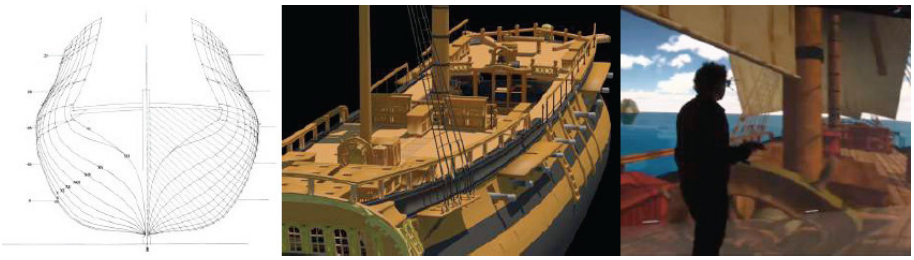


Figure 4. Digital reconstitution of "Le Boullongne". From architectural plans to virtual reality implementation.

Touching and interacting with inaccessible cultural heritage Valérie Gouranton and Bruno Arnaldi

Sense of touch provides a particular access to our environment, enabling a tangible relation with it. In the particular use case of cultural heritage, touching the past, apart from being a universal dream, can provide essential information to analyze, understand, or restore artifacts. However, archaeological objects cannot always offer a tangible access, either because they have been destroyed or too damaged, or because they are part of a larger assembly. In other cases, it is the context of use that has become inaccessible, as it is related to an extinct activity. In [15] we proposed a workflow based on a combination of computed tomography, 3D images, and 3D printing to provide concrete access to cultural heritage, and we illustrate this workflow in different contexts of inaccessibility. These technologies are already used in cultural heritage, but seldom combined, and mostly for exceptional artifacts. We proposed to combine these technologies in case studies corresponding to relevant archaeological situations.

This work was done in collaboration with Théophane Nicolas (INRAP), Ronan Gaugne (Univ. Rennes 1), Cédric Tavernier (Image ET) and Quentin Petit (CNRS).

3D reconstruction of the loyola sugar plantation and virtual reality applications Jean-Baptiste Barreau, Valérie Gouranton

Discovered in 1988, the Loyola sugar plantation, owned by the Jesuits in French Guiana, is a major plantation of colonial history and slavery. Ongoing archaeological excavations have uncovered the Jesuit's house and the outbuildings usually associated with a plantation such as a chapel and its cemetery, a blacksmith shop, a pottery, the remains of the entire sugar production (a windmill, a boiler and a dryer), coffee and indigo warehouses etc. Based on our findings and our network with 3D graphic designers and researchers in virtual reality, a 3D restitution integrated within a virtual reality platform was initiated to develop a better understanding of the plantation and its surrounding landscape. A specific work on the interactive changes of sunlight and animal sounds aimed to reconstruct a coherent evolution during one day of the site's environment [21].

This work was done in collaboration with Quentin Petit (CNRS),Yann Bernard (CReAAH), Reginald Auger (Laval University, Canada), Yannick Le Roux (Laval University, French Guiana) Ronan Gaugne (IMMERSIA), and Cédric Tavernier (Image ET).

7.2. Physically-Based Simulation and Multisensory Feedback

7.2.1. Interactive Physically-Based Simulation

Aggregate constraints for virtual manipulation with soft fingers, Maud Marchal, Anthony Talvas

Figure 5. Interaction with deformable fingers generates many interconnected contact points which are expensive to solve with friction. Our approach aggregates contact constraints per phalanx with torsional friction. The subsequent increase in performance allows real time dexterous manipulation of virtual objects using soft fingers.

Interactive dexterous manipulation of virtual objects remains a complex challenge that requires both appropriate hand models and accurate physically-based simulation of interactions. In [16], we proposed an approach based on novel aggregate constraints for simulating dexterous grasping using soft fingers. Our approach aims at improving the computation of contact mechanics when many contact points are involved, by aggregating the multiple contact constraints into a minimal set of constraints. We also introduced a method for non-uniform pressure distribution over the contact surface, to adapt the response when touching sharp edges. We used the Coulomb-Contensou friction model to efficiently simulate tangential and torsional friction. We showed through different use cases that our aggregate constraint formulation is well-suited for simulating interactively dexterous manipulation of virtual objects through soft fingers, and efficiently reduces the computation time of constraint solving. This work was done in collaboration with Christian Duriez (Inria team DEFROST) and Miguel Otaduy (Univ. Rey Juan Carlos, Madrid, Spain).

7.2.2. Multimodal Feedback

Elastic-Arm: Human-scale passive feedback for augmenting interaction and perception in virtual environments Merwan Achibet, Adrien Girard, Maud Marchal, Anatole Lécuyer

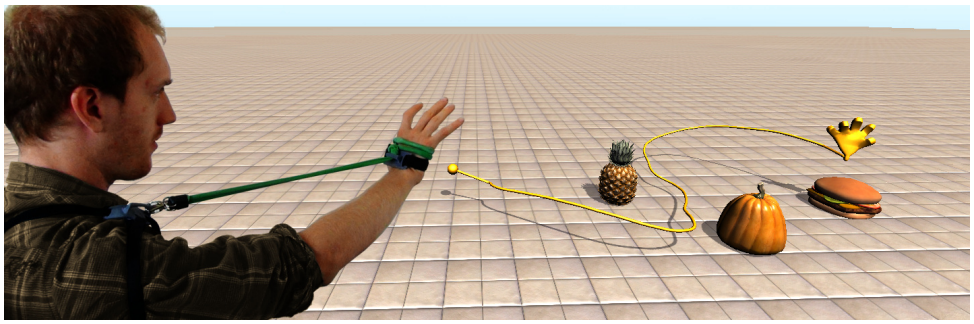


Figure 6. The Elastic-Arm is a body-mounted armature that provides egocentric passive haptic feedback. It presents an alternative to more complex active haptic devices that are generally less adapted to large immersive environments. In this example, a user performs a selection task by stretching his virtual arm using a combination of the Bubble and Go-Go techniques reimplemented with our system.

Haptic feedback is known to improve 3D interaction in virtual environments but current haptic interfaces remain complex and tailored to desktop interaction. In [18], we introduced the *ElasticArm*, a novel approach for incorporating haptic feedback in immersive virtual environments in a simple and cost-effective way. The Elastic-Arm is based on a body-mounted elastic armature that links the user's hand to her shoulder. As a result, a progressive resistance force is perceived when extending the arm. This haptic feedback can be incorporated with various 3D interaction techniques and we illustrate the possibilities offered by our system through several use cases based on well-known examples such as the *Bubble* technique, *Redirected Touching*, and pseudo-haptics. These illustrative use cases provide users with haptic feedback during selection and navigation tasks but they also enhance their perception of the virtual environment. Taken together, these examples suggest that the Elastic-Arm can be transposed in numerous applications and with various 3D interaction metaphors in which a mobile haptic feedback can be beneficial. It could also pave the way for the design of new interaction techniques based on human-scale egocentric haptic feedback.

Visual vibrations to simulate taps on different materials Maud Marchal, Anatole Lécuyer

In [40], we presented a haptic visualization technique for conveying material type through visual feedback, expressed as visible decaying sinusoidal vibration resulting from tapping an object. The technique employs cartoon-inspired visual effects and modulates the scale of the vibration to comply with visual perception. The results of a user study showed that participants could successfully perceive three types of material (rubber, wood, and aluminum) using our novel visual effect. This work was done in collaboration with Taku Hachisu and Hiroyuki Kajimoto (Univ. of Electro Communication, Tokyo, Japan).

7.2.3. GPU-based Collision Detection in Virtual Environments

GPU Ray-Traced Collision Detection: Fine Pipeline Reorganization François Lehericey, Valérie Gouranton, Bruno Arnaldi

Ray-tracing algorithms can be used to render a virtual scene and to detect collisions between objects. Numerous ray-tracing algorithms have been proposed which use data structures optimized for specific cases (rigid objects, deformable objects, etc.). Some solutions try to optimize performance by combining several algorithms to use the most efficient algorithm for each ray. In [31], we presented a ray-traced collision detection pipeline that improves the performance on a graphic processing unit (GPU) when several ray-tracing algorithms are used.

When combining several ray-tracing algorithms on a GPU, a well-known drawback is thread divergence among work-groups that can cause loss of performance by causing idle threads. We avoid branch divergence

by dividing the ray tracing into three steps with appended buffers in between. We also show that prediction can be used to avoid unnecessary synchronizations between the CPU and GPU. Applied to a narrow-phase collision detection algorithm, results show an improvement of performance up to 2.7 times.

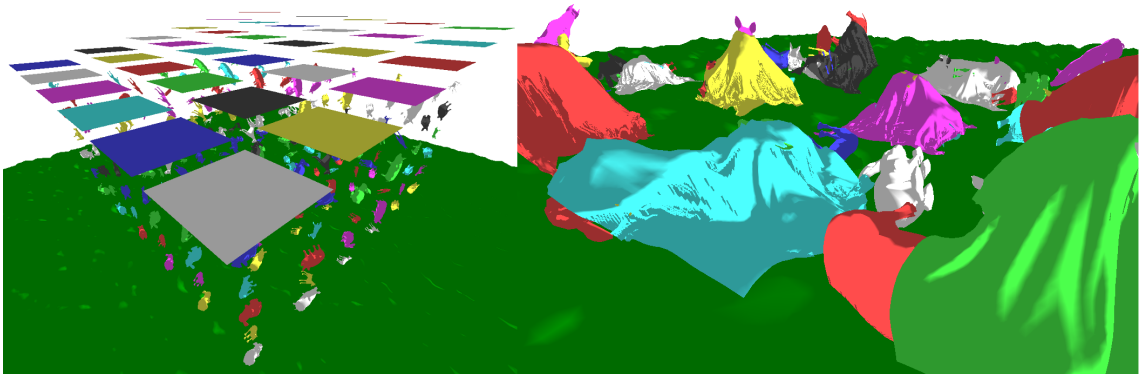


Figure 7. 216 concave objects fall on an irregular ground and 36 deformable sheets fall over them [31].

GPU Ray-Traced Collision Detection for Cloth Simulation François Lehericey, Valérie Gouranton, Bruno Arnaldi

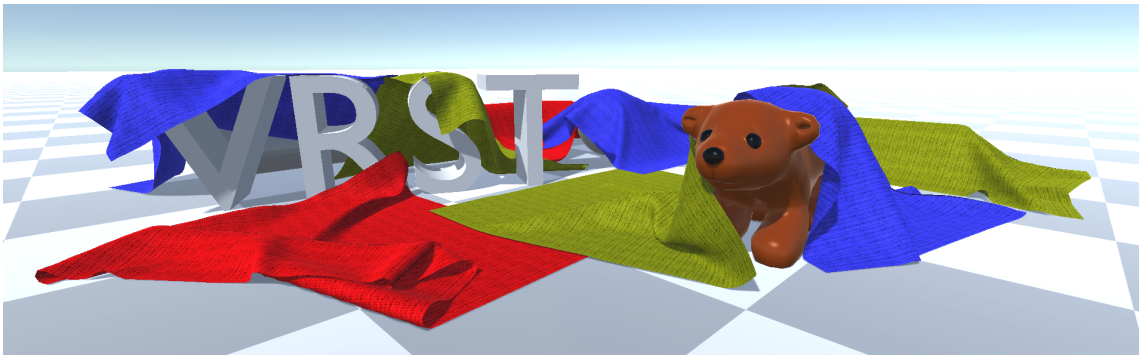


Figure 8. Our method can perform collision detection between clothes and handle self collision detection [30].

In [30], we proposed a method to perform collision detection with cloths with ray-tracing at an interactive frame-rate. Our method is able to perform collision detection between cloths and volumetric objects (rigid or deformable) as well as collision detection between cloths (including auto-collision). Our method casts rays between objects to perform collision detection, and an inversion-handling algorithm is introduced to correct errors introduced by discrete simulations. GPU computing is used to improve the performances by parallelizing the ray-tracing. Our implementation handles scenes containing deformable objects at an interactive frame-rate, with collision detection lasting a few milliseconds.

7.2.4. Medical Applications

Real-time tracking of deformable targets in 3D ultrasound images Maud Marchal

In [35], [36], we presented a novel approach for tracking a deformable anatomical target within 3D ultrasound volumes. Our method is able to estimate deformations caused by the physiological motions of the patient. The displacements of moving structures are estimated from an intensity-based approach combined with a physically-based model and has therefore the advantage to be less sensitive to the image noise. Furthermore, our method does not use any fiducial marker and has real-time capabilities. The accuracy of our method is evaluated on real data acquired from an organic phantom. The validation is performed on different types of motions comprising rigid and non-rigid motions. Thus, our approach opens novel possibilities for computer-assisted interventions where deformable organs are involved.

Our approach was also evaluated on the MICCAI CLUST'15 challenge 3D database. We achieved a mean tracking error of 1.78 mm with an average computation time of 350 ms per frame, ranking our method first during the on-site challenge [34]. This work was done in collaboration with Lucas Royer, Anthony Le Bras and Guillaume Dardenne (IRT bcom), and Alexandre Krupa (Inria team LAGADIC).

Statistical study of parameters for deep brain stimulation automatic pre-operative planning of electrodes trajectories Maud Marchal

Automatic methods for pre-operative trajectory planning of electrodes in Deep Brain Stimulation are usually based on the search for a path that resolves a set of surgical constraints to propose an optimal trajectory. In [13], we studied the use of parameters based on real trajectories of surgeons. For that purpose we firstly retrieve the actual weighting factors used by neurosurgeons thanks to a retrospective study, secondly we compare the results from two different hospitals to evaluate their similarity, and thirdly we compare these trends to the weighting factors usually empirically set in most current approaches. We proposed two approaches, one based on a stochastic sampling and the other on an exhaustive search. In each case, we get a sample of combinations of weighting factors along with a measure of their quality, i.e. the similarity between the optimal trajectory they lead to and the trajectory manually planned by the surgeon as a reference. Then visual and statistical analysis are performed on the number of occurrences and on the rank means. We performed our study on 56 retrospective cases from two different hospitals.We could observe a trend of the occurrence of each weight on the number of occurrences. We also proved that each weight had a significant influence on the ranking. Additionally, we observed no influence of the medical center parameters, suggesting that the trends were comparable in both hospitals. Finally, the obtained trends were confronted to the usual weights chosen by the community, showing some common points but also some discrepancies. These results tend to show a predominance of the choice of a trajectory close to a standard direction. Secondly, the avoidance of the vessels or sulci seems to be sought in the surroundings of the standard position. The avoidance of the ventricles seem to be less predominant, but this could be due to the already reasonable distance between the standard direction and the ventricles. The similarity of results between two medical centers tend to show that it is not an exceptional practice. This work was done in collaboration with Caroline Essert and Antonio Capobianco (Univ. Strasbourg), Claire Haegelen and Pierre Jannin (LTSI, Rennes), Sara Fernandez-Vidal, Carine Karachi and Eric Bardinet (Institut du Cerveau et de la Moëlle Epinière, Paris).

7.3. Collaborative Virtual Environments

Asymetric Remote Collaboration in Mixed Reality: Awareness and Navigation Morgan Le Chénéchal, Valérie Gouranton and Bruno Arnaldi

We first focused on the lack of mutual awareness that may appear in many situations and we evaluated different ways to present the distant user and his actions in the Virtual Environment (VE) in order to understand his perception and cognitive process. We focused on a common case consisting in estimating accurately the time at which a distant user analyzed the meaning of a remotely pointed object. Amongst others, our experimental results presented at CTS [28], show that expertise of the users influences on how they estimate the distant activity and the type of applied strategies.

Then, in a similar asymmetric setup, we proposed a demo at IEEE VR to deal with real estate business. In this context, it is quite difficult for estate agents to make customers understand the potential and the volumes of free spaces. The demo aimed to solve these issues based on a laying out scenario in which a seller and a

customer collaborate. As the roles of both users are different, we proposed an asymmetric collaboration where the two users do not use the same interaction setup and do not benefit from the same interaction capabilities.

Last, we focused on a remote collaborative maintenance scenario in which a remote expert helps an operator in performing a physical task 9 . Our system is based on a VR setup for the remote expert in order to virtually co-locates him in the real workspace, and an AR interface for the display of the helping gestures to the agent. In a preliminary user study, we evaluated the performance of our system in a navigation task, and we presented results at ICAT-EGVE [29].

This work was done in collaboration with Thierry Duval (Lab-STICC) and Jérome Royan (IRT B-Com).

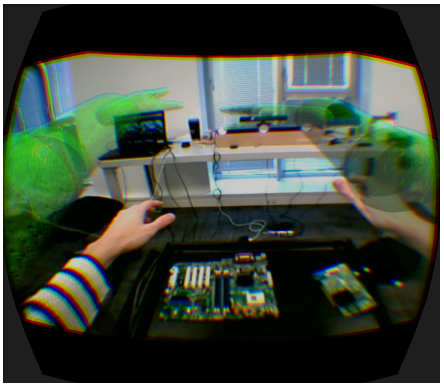


Figure 9. Remote collaborative maintenance using mixed reality.

High-Level Components for Developing Collaborative and Interactive Virtual Environments Rozenn Bouville, Valérie Gouranton, Thomas Boggini, Florian Nouviale and Bruno Arnaldi

We proposed a framework called #FIVE (Framework for Interactive Virtual Environments) for the development of interactive and collaborative virtual environments [22]. It has been developed for an easier and a faster design and development of virtual reality applications. It was designed with a constant focus on re-usability with as few hypotheses as possible on the final application in which it could be used. Whatever the chosen implementation for the Virtual Environment (VE), #FIVE : (1) provides a toolkit that eases the declaration of possible actions and behaviours of objects in the VE, (2) provides a toolkit that facilitates the setting and the management of collaborative interactions in a VE, (3) is compliant with distribution of the VE on different setups and (4) proposes guidelines to efficiently create a collabo-rative and interactive VE. It is composed of several modules, among them, two core modules : the relation engine and the collaborative interaction engine. On the one hand, the relation engine manages the relations between the objects of the environment. On the other hand, the collaborative interaction engine manages how users can collaboratively control objects. The modules that compose the #FIVE framework can be used either independently or simultaneously , depending on the requirements of the application. They can also communicate and work with other modules thanks to an API. For instance, a scenario engine can be plugged to any or both of the #FIVE modules if the application is scenario-based. #FIVE has already been used in VR applications by several members of our team (see section 6.5). The feedbacks are rather positive and we intend to further develop #FIVE with additional functionalities, notably by extending it to the control of avatars whether they are controlled by a user or by the system.

High-Level Components for Developing Collaborative Scenarios Guillaume Claude, Valérie Gouranton and Bruno Arnaldi

We were interested in the description of activities of actors in Collaborative Virtual Environments for Training to team working on procedures. We have proposed #SEVEN, a model for the description of procedures as

Collaborative Virtuel Environments Scenarios (see also section 6.6). In [25] we have demonstrated the abilities of this model to be adapted to a wide range of use cases. We showed that it can adapt its abstraction level to the required guidance level and describe more or less complex unfolding of events. In [24] we have provided a novel approach to the distribution of the actions between the actors of the simulation by using an action filtering model in conjunction with a reactive team model. The action filtering model uses data about the actors such as their abilities or their rights. Our reactive team model can be used to define relationships between the team members and the effects of inner rules of the team upon the involvement of the actors in the procedure. To our knowledge, our solution is the closest to the existing models proposed by the social science domain known as role theory. Our work has been applied to several domains, including the training of scrub nurses to neurosurgery procedures 10 .

Figure 10. The #FIVE and #SEVEN models used in the S3PM project to provide an interactive environment and define collaborative scenarios and handle dynamic team structures in a surgical context.

7.4. Brain-Computer Interfaces

7.4.1. Novel Usages of BCI

Mind-Window: Real-Time Brain Activity Visualization Using Tablet-Based Augmented Reality and EEG for Single or Multiple Users, Anatole Lécuyer, Jonathan Mercier, Maud Marchal

Figure 11. Our novel "Mind-Window" approach enables one or multiple users to visualize the brain activity of a person in real-time by using tablets and augmented reality. It proposes to see through the tablet a virtual brain model "as if the skull is transparent". The display of the virtual brain is updated in real-time according to the real brain activity of the person which is measured thanks to an EEG headset.

We introduced a novel approach, called the “Mind-Window”, for real-time visualization of brain activity [33]. The Mind-Window enables one or multiple users to visualize the brain activity of another person as if his/her skull was transparent. Our approach relies on the use of multiple tablet PCs that the observers can move around the head of the observed person wearing an electroencephalography cap (EEG). A 3D virtual brain model is superimposed to the head of the observed person using augmented reality by tracking a 3D marker placed on top the head. The EEG cap records the electrical fields emitted by the brain, and they are processed in real-time to update the display of the virtual brain model. Several visualization techniques are proposed such as an interactive cutting plane which can be manipulated with touch-based inputs on the tablet. The Mind-Window approach could be used for medical applications, e.g. by providing a simple way for physicians to diagnose and observe brain activity of patients. Teachers could also use our system to teach brain anatomy/activity and EEG features, e.g., electrodes localization, electrical patterns, etc. Finally, video conferences or video games could be “brain-augmented”, making use of the Mind-Window for entertainment purposes.

B-C-Invisibility Power: Optical Camouflage Based on Mental Activity in Augmented Reality, Anatole Lécuyer, Jonathan Mercier, Maud Marchal

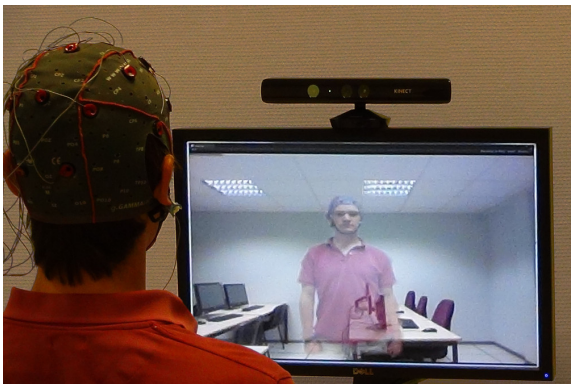


Figure 12. The "B-C-Invisibility power” enables users to become virtually invisible by performing mental tasks. Brain signals are extracted using EEG electrodes and analyzed within the BCI.

In the context of the ANR project HOMO-TEXTILUS which focuses on the design of novel "smart clothes", we introduced a kind of "invisibility cloak": an interactive approach for using Brain-Computer Interfaces for controlling optical camouflage called "B-C-Invisibility power”. We proposed to combine augmented reality and BCI technologies to design a system which somehow provides the "power of becoming invisible” [32]. Our optical camouflage is obtained on a PC monitor combined with an optical tracking system. A cut out image of the user is computed from a live video stream and superimposed to the prerecorded background image using a transparency effect. The transparency level is controlled by the output of a BCI, making the user able to control her invisibility directly with mental activity. The mental task required to increase/decrease the invisibility is related to a concentration/relaxation state. Results from a preliminary study based on a simple video-game inspired by the Harry Potter universe could notably show that, compared to a standard control made with a keyboard, controlling the optical camouflage directly with the BCI could enhance the user experience and the feeling of “having a super-power”.

7.4.2. BCI Methodology and Techniques

A methodological framework for applications combining BCI and videogames, Anatole Lécuyer

We have proposed a user-centered methodological framework [41] to guide design and evaluation of applications based on Brain-Computer Interface (BCI). Our framework is based on the contributions of ergonomics

to ensure that these applications are well suited for end-users. It provides methods, criteria and metrics to perform the phases of the human-centered design process aiming to understand the context of use, specify the user needs and evaluate the solutions in order to define design choices. Several ergonomic methods (e.g., interviews, longitudinal studies, user based testing), objective metrics (e.g., task success, number of errors) and subjective metrics (e.g., mark assigned to an item) are suggested to define and measure the usefulness, usability, acceptability, hedonic qualities, appealingness, emotions related to user experience, immersion and presence to be respected. The benefits and contributions of our user centred framework for the ergonomic design of videogames based on BCI were also discussed.

This work was done in collaboration with Fabien Lotte (Inria team POTIOC).

Feasibility and specificity of simultaneous EEG and fMRI, Marsel Mano, Lorraine Perronnet, Jussi Lindgren, Anatole Lécuyer

In the field of fMRI, Arterial Spin Labeling (ASL) imaging relies on control and label radio-frequency pulses. This generates alternate gradient patterns as well as higher specific absorption rate (SAR). To date, only a few studies have addressed the issue of connecting EEG signal to ASL perfusion. Furthermore, previous studies have shown reduced blood-oxygen-level dependent (BOLD) signal-to-noise ratio (SNR) in the presence of EEG. ASL being a low SNR technique, the aim of this study was to assess ASL-EEG at 3T in terms of safety as well as EEG and magnetic resonance signal quality. Our experimental results show that ASL-EEG can be safely performed [20] [38]. Standard ASL acquisitions generated more than 2.5-fold SAR increase compared to a standard BOLD echo planar imaging sequence. This corresponded to up to 4°C temperature increase on the bundle, yet not exceeding 36°C. Gradient artifact correction of the EEG signal by average artifact subtraction was generally good for BOLD-EEG and ASL-EEG. However, residual gradient artifacts affecting 1% of the pulsed ASL-EEG data have to be considered. Further research is needed to understand the artifact variability and to develop an appropriate correction strategy. No residual artifacts were observed for alternating control and label pulses ASL-EEG. Neither a change of the number of reference volumes for artifact subtraction nor an independent component analysis could help tackle this gradient artifact correction issue. Regarding magnetic resonance imaging, a 20% loss in SNR was observed when compared to acquisitions performed without EEG. Taken together our results suggest that EEG and ASL MRI can be simultanously combined for the purpose of real-time experiments which could for instance be envisionned in our HEMISFER project.

This work was done in collaboration with VISAGES team.

LAGADIC Project-Team

7. New Results

7.1. Visual tracking

7.1.1. Object detection

Participant: Eric Marchand.

We addressed the challenge of detecting and localizing a poorly textured known object, by initially estimating its complete 3D pose in a video sequence [45]. Our solution relies on the 3D model of the object and synthetic views. The full pose estimation process is then based on foreground/background segmentation and on an efficient probabilistic edge-based matching and alignment procedure with the set of synthetic views, classified through an unsupervised learning phase. Our study focuses on space robotics applications and the method has been tested on both synthetic and real images, showing its efficiency and convenience, with reasonable computational costs.

7.1.2. Registration of multimodal images

Participant: Eric Marchand.

This study has been realized in collaboration with Brahim Tamadazte and Nicolas Andreff from Femto-ST, Besançon. Following our developments in visual tracking and visual servoing from the mutual information [3], it concerned mutual information-based registration of white light images vs. fluorescence images for micro-robotic laser microphonosurgery of the vocal folds. Nelder-Mead Simplex for nonlinear optimization has been used to minimize the cost-function [43].

7.1.3. Pose estimation from RGB-D sensor

Participant: Eric Marchand.

RGB-D sensors have become in recent years a product of easy access to general users. They provide both a color image and a depth image of the scene and, besides being used for object modeling, they can also offer important cues for object detection and tracking in real-time. In this context, the work presented in this paper investigates the use of consumer RGB-D sensors for object detection and pose estimation from natural features. Two methods based on depth-assisted rectification are proposed, which transform features extracted from the color image to a canonical view using depth data in order to obtain a representation invariant to rotation, scale and perspective distortions. While one method is suitable for textured objects, either planar or non-planar, the other method focuses on texture-less planar objects [18]

7.1.4. 3D localization for airplane landing

Participants: Noël Mériaux, François Chaumette, Patrick Rives, Eric Marchand.

This study is realized in the scope of the ANR VisioLand project (see Section 9.2.2 . In a first step, we have considered and adapted our model-based tracker [2] to localize the aircraft with respect to the airport surroundings. Satisfactory results have been obtained from real image sequences provided by Airbus. In a second step, we have started to perform this localization from a set of keyframe images corresponding to the landing trajectory.

7.2. Visual servoing

7.2.1. Histogram-based visual servoing

Participants: Quentin Bateux, Eric Marchand.

Classically visual servoing considers the regulation in the image of a set of visual features (usually geometric features). Direct visual servoing schemes, such as photometric visual servoing, have been introduced in order to consider every pixel of the image as a primary source of information and thus avoid the extraction and the tracking of such geometric features. This year, we proposed a method to extend these works by using a global descriptor, namely intensity histograms, on the whole or multiple sub-sets of the images in order to achieve control of a 6 degrees of freedom (DoF) robot [30][53].

7.2.2. Photometric moment-based visual servoing

Participants: Manikandan Bakthavatchalam, François Chaumette.

This work also belongs to the class of direct visual servoing. Its goal was to use photometric moments as visual features in order to increase the convergence domain of this approach by reducing the non linearity of the control problem. In order to cope with appearance and disappearance of some parts of the environment during the camera motion, a spatial weight has been introduced in the definition of photometric moments. Thanks to a particular design of this weight, the analytical form of the interaction matrix has been obtained, from which it was possible to select a set of moment combinations to control all the six degrees of freedom of the system. Satisfactory experimental results have been obtained [29][8], even if the loss of invariance properties makes the optimal design of visual features still an open problem.

7.2.3. Model predictive visual servoing

Participants: Nicolas Cazy, Paolo Robuffo Giordano, François Chaumette.

The goal of this work is to exploit Model Predictive Control (MPC) techniques for dealing in a robust way with loss of features during a IBVS task. The work [31] provides an experimental validation of different correction schemes able to cope with loss of features due to occlusions of limited camera field of view. The reported results show the effectiveness of the proposed techniques during the servoing of four point features.

7.2.4. Nanomanipulation

Participants: Le Cui, Eric Marchand.

Following our work related to scanning electron micro- scope (SEM) calibration [12] we considered the control of a micro robot using a direct photometric visual servoing that uses only the pure image information as a visual feature, instead of using classic geometric features such as points or lines. However, in micro-scale, using only image intensity as a visual feature performs unsatisfactorily in cases where the photometric variation is low, such as motions along vision sensor's focal axis under a high magnification. In order to improve the performance and accuracy in those cases, an approach using hybrid visual features is proposed in this paper. Image gradient is employed as a visual feature on z axis while image intensity is used on the other 5 DoFs to control the motion. A 6-DoF micro-positioning task is accomplished by this hybrid visual servoing scheme [34].

We also considered a full scale autofocus approach for SEM [35]. The optimal focus (in-focus) position of the microscope is achieved by maximizing the image sharpness using a vision-based closed-loop control scheme. An iterative optimization algorithm has been designed using the sharpness score derived from image gradient information. The proposed method has been implemented and validated using a tungsten gun SEM at various experimental conditions like varying raster scan speed, magnification at real-time.

7.2.5. Audio-based control

Participants: Aly Magassouba, François Chaumette.

This study is not concerned with visual servoing, but to the application of the same principle of sensor-based control to audio sensors. It is made in collaboration with Nancy Bertin from Panama group at Irisa, Inria Rennes-Bretagne Atlantique. In a first step, we have determined the analytical form of the interaction matrix of audio features based on the time difference of arrival on two microphones. From this modeling step, we have determined the different virtual linkages that can be realized in function of the number and configuration of sources [41]. First experimental results using two microphones monunted on the Pioneer mobile robot (see Section 6.9) have been recently obtained.

7.3. Visual navigation of mobile robots

7.3.1. Visual navigation from straight lines

Participants: Suman Raj Bista, Paolo Robuffo Giordano, François Chaumette.

This study is concerned with visual autonomous navigation in indoor environments. As in our previous works concerning navigation outdoors [4], the approach is based on a topological localization of the current image with respect to a set of keyframe images, but the visual features used for this localisation as well as for the visual servoing is not based on points of interest, but straight lines that are more common indoors. Satisfactory experimental results have been obtained using the Pioneer mobile robot (see Section 6.9) [23].

7.3.2. Autonomous navigation of a wheelchair and social navigation

Participants: Vishnu Karakkat Narayanan, François Pasteau, Marie Babel.

Navigating within an unknown indoor environment using an electric wheelchair is a challenging task, especially if the user suffers from severe disabilities. We presented in [22] a framework for vision-based autonomous indoor navigation in an electric wheelchair capable of following corridors, and passing through open doorways using a single doorpost. The designed control schemes have been implemented onto a robotized wheelchair and experimental results show the robust behaviour of the designed system.

We then introduced in [40] a task-based control law which can serve as a low-level system for equitably joining interacting groups, while confirming to social conventions. The system uses the position and orientation of the participating humans with respect to a rigid sensor frame in order to control the translational and rotational velocity of a wheelchair so that the robot positions itself aptly at the meeting point

7.3.3. Semi-autonomous control of a wheelchair for navigation assistance

Participants: Vishnu Karakkat Narayanan, François Pasteau, Marie Babel.

To address the wheelchair driving assistance issue, we proposed in [56][28] a unified shared control framework able to smoothly correct the trajectory of the electrical wheelchair. The system integrates the manual control with sensor-based constraints by means of a dedicated optimization strategy. The resulting low-complex and low-cost embedded system is easily plugged onto on-the-shelf wheelchairs.

The robotic solution has been then validated through clinical trials that have been conducted within the Rehabilitation Center of Pôle Saint Hélier (France) with 25 volunteering patients presenting different disabling neuro-pathologies. This assistive tool is shown to be intuitive and robust as it respects the user intention, it does not alter perception while reducing the number of collisions in case of hazardous maneuvers or in crowded environment [27].

7.4. 3D Scene Mapping

7.4.1. 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 [50] we have addressed the active estimation of the 3D structure of an observed planar scene by comparing three different techniques: a homography decomposition (a well-established method taken as a baseline), a least-square fitting of a reconstructed 3D point cloud, and a direct estimation based on the observation of a set of discrete image moments made of a collection of image points belonging to the observed plane. The experimental results confirmed the importance of actively controlling the camera motion in order to obtained a faster convergence for the estimation error, as well as the superiority of the third method based on the machinery of image moments for what concerns robustness against noise and outliers. In [51] the active estimation scheme has been improved by considering a set of features invariant to camera rotations. This

way, the dynamics of the structure estimation becomes independent of the camera angular velocity whose measurement is, thus, no longer required for implementing the active SfM scheme. Finally, in [46] the issue of determining online the 'best' combination of image moments for reconstructing the scene structure has been considered. By defining a new set of weighted moments as a weighted sum of traditional image moments, it is indeed possible to optimize for the weights online during the camera motion. The SfM scheme then automatically selects online the best combination of image moments to be used as measurements as a function of the current scene.

7.4.2. Scene Registration based on Planar Patches

Participants: Eduardo Fernandez Moral, Patrick Rives.

Scene registration consists of estimating the relative pose of a camera with respect to a scene previously observed. This problem is ubiquitous in robot localization and navigation. We propose a probabilistic framework to improve the accuracy and efficiency of a previous solution for structure registration based on planar representation. Our solution consists of matching graphs where the nodes represent planar patches and the edges describe geometric relationships. The maximum likelihood estimation of the registration is estimated by computing the graph similarity from a series of geometric properties (areas, angles, proximity, etc..) to maximize the global consistency of the graph. Our technique has been validated on different RGB-D sequences, both perspective and spherical [14].

7.4.3. Robust RGB-D Image Registration

Participants: Tawsif Gokhool, Renato José Martins, Patrick Rives.

Estimating dense 3D maps from stereo sequences remains a challenging task where building compact and accurate scene models is relevant for a number of tasks, from localization and mapping to scene rendering [20], [10] . In this context, this work deals with generating complete geometric and photometric "minimal" model of indoor/outdoor large-scale scenes, which are stored within a sparse set of spherical images to asset photo-geometric consistence of the scene from multiple points-of-views . To this end, a probabilistic data association framework for outlier rejection is formulated, enhanced with the notion of landmark stability over time. The approach was evaluated within the frameworks of image registration, localization and mapping, demonstrating higher accuracy and larger convergence domains over different datasets [39].

7.4.4. Accurate RGB-D Keyframe Representation of 3D Maps

Participants: Renato José Martins, Eduardo Fernandez Moral, Patrick Rives.

Keyframe-based maps are a standard solution to produce a compact map representation from a continuous sequence of images, with applications in robot localization, 3D reconstruction and place recognition. We have present a approach to improve keyframe-based maps of RGB-D images based on two main filtering stages: a regularization phase in which each depth image is corrected considering both geometric and photometric image constraints (planar and superpixel segmentation); and a fusion stage in which the information of nearby frames (temporal continuity of the sequence) is merged (using a probabilistic framework) to improve the accuracy and reduce the uncertainty of the resulting keyframes. As a result, more compact maps (with less keyframes) are created. We have validated our approach with different kind of RGB-D data including both indoor and outdoor sequences, and spherical and perspective sensors, demonstrating that our approach compares and outperforms the state-of-the-art [42].

7.4.5. Semantic Representation For Navigation In Large-Scale Environments

Participants: Romain Drouilly, Patrick Rives.

Autonomous navigation is one of the most challenging problem to address to allow robots to evolve in our everyday environments. Map-based navigation has been studied for a long time and researches have produced a great variety of approaches to model the world. However, semantic information has only recently been taken into account in those models to improve robot efficiency.

Mimicking human navigation is a challenging goal for autonomous robots. This requires to explicitly take into account not only geometric representation but also high-level interpretation of the environment [9]. We propose a novel approach demonstrating the capability to infer a route in a global map by using semantics. Our approach relies on an object-based representation of the world automatically built by robots from spherical images. In addition, we propose a new approach to specify paths in terms of high-level robot actions. This path description provides robots with the ability to interact with humans in an intuitive way. We perform experiments on simulated and real-world data, demonstrating the ability of our approach to deal with complex large-scale outdoor environments whilst dealing with labelling errors [37].

Mapping evolving environments requires an update mechanism to efficiently deal with dynamic objects. In this context, we propose a new approach to update maps pertaining to large-scale dynamic environments with semantics. While previous works mainly rely on large amount of observations, the proposed framework is able to build a stable representation with only two observations of the environment. To do this, scene understanding is used to detect dynamic objects and to recover the labels of the occluded parts of the scene through an inference process which takes into account both spatial context and a class occlusion model. Our method was evaluated on a database acquired at two different times with an interval of three years in a large dynamic outdoor environment. The results point out the ability to retrieve the hidden classes with a precision score of 0.98. The performances in term of localisation are also improved [36].

7.5. Control of single and multiple Unmanned Aerial Vehicles

7.5.1. Single UAV

Participant: 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.

Despite these clear advantages, a clear shortcoming of the quadrotor design lies in its inherent underactuation (only 4 actuated propellers for the 6 dofs of the quadrotor pose). This underactuation limits the quadrotor flying ability in free or cluttered space and, furthermore, it also degrades the possibility of interacting with the environment by exerting desired forces in arbitrary directions. In [24], a novel design for a quadrotor UAV with tilting propellers which is able to overcome these limitations has been presented and experimentally validated. Indeed, the additional set of 4 control inputs actuating the propeller tilting angles can be shown to yield full actuation to the quadrotor position/orientation in space, thus allowing it to behave as a fully-actuated flying vehicle and to overcome the aforementioned underactuation problem.

Furthermore, the issue of estimating online the UAV self-motion from vision has been considered. To this end, a novel nonlinear estimation scheme able to recover the metric UAV linear velocity from the *scaled* one obtained from the decomposition of the optical flow has been proposed in [15]. The observability conditions (in terms of persistency of excitation) needed to ensure a converging estimation have also been studied. The reported experimental results confirmed the effectiveness of the estimation scheme in recovering a reliable and accurate estimation of the UAV self-motion (linear and angular velocities) in realistic conditions.

This work has been realized in collaboration with the Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

7.5.2. Collective control of multiple UAVs

Participants: Fabrizio Schiano, 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. However, connectivity alone is not sufficient to perform certain tasks when only *relative sensing* is used. For these systems, the concept of *rigidity* provides the correct framework for defining an appropriate sensing and communication topology architecture. Rigidity is a combinatorial theory for characterizing the "stiffness" or "flexibility" of structures formed by rigid bodies connected by flexible linkages or hinges. In a broader context, rigidity turns out to be an important architectural property of many multi-agent systems when a common inertial reference frame is unavailable. Applications that rely on sensor fusion for localization, exploration, mapping and cooperative tracking of a target, all can benefit from notions in rigidity theory. The concept of rigidity, therefore, provides the theoretical foundation for approaching decentralized solutions to the aforementioned problems using distance measurement sensors, and thus establishing an appropriate framework for relating system level architectural requirements to the sensing and communication capabilities of the system.

In [26], a decentralized gradient-based rigidity maintenance action for a group of quadrotor UAVs has been proposed and tested in real experimental conditions. By starting in a rigid configuration, the group of UAVs is able to estimate their relative position from sole relative distance measurements, and then use these estimated relative positions in a control action able to preserve rigidity of the whole formation despite presence of sensor limitations (maximum range and line-of-sight occlusions), possible collisions with obstacles and inter-robot collisions. Furthermore, in [52] the novel case of *bearing rigidity* for directed graphs has been considered: here, rather than distances the measurements are the 3D bearing vectors expressed in the local body-frame of each agent. The theory has been developed for the case of planar agents in $SE(2)$ and a 'scale-free' bearing controller has been proposed, able to steer the robot group towards a desired bearing formation.

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

7.5.3. Cooperative localization using interval analysis

Participants: Vincent Drevelle, Ide Flore Kenmogne Fokam.

In the context of multi-robot fleets, cooperative localization consists in gaining better position estimate through measurements and data exchange with neighboring robots. Positioning integrity (i.e., providing reliable position uncertainty information) is also a key point for mission-critical tasks, like collision avoidance. The goal of this work is to compute position uncertainty volumes for each robot of the fleet, using a decentralized method (i.e using only local communication with the neighbors). The problem is addressed in a bounder-error framework, with interval analysis and constraint propagation methods. These methods enable to provide guaranteed position error bounds, assuming bounded-error measurements. They are not affected by over-convergence due to data incest, which makes them a well sound framework for decentralized estimation. Encouraging results have already been obtained for multi-robot underwater positioning with acoustical range measurements. Ongoing work focuses on cooperative localization in a multi-UAV fleet with image-based measurements (bearings).

7.6. Medical robotics

7.6.1. Non-rigid target tracking in ultrasound images combining dense information and physically-based model

Participants: Lucas Royer, Alexandre Krupa.

This study concerns the real-time tracking of deformable targets within a sequence of ultrasound (US) images. The proposed approach combines dense information with a physically-based model and has therefore the advantage of not using any fiducial marker. The physical model is represented by a mass-spring damper system driven by external and internal forces. The external forces are obtained by maximizing an image similarity metric between a reference target and the deformed target along the time. The internal forces of the mass-spring damper system constrain the deformation to be physically plausible and therefore efficiently reduce the sensitivity to the speckle noise. This approach was first validated from simulated and real sequences of 2D US images [49]. It was then extended for deformable target tracking in a sequence of 3D ultrasound volumes and tested on a robotic setup used to apply deformation on an organic phantom [48]. The performance of this deformable 3D target tracking approach was evaluated with visual assessment combined with robotic odometry ground truth. This method was also tested and compared with respect to state-of-the-art techniques by using 3D image databases provided by MICCAI CLUST'14 and CLUST'15 challenges [47] (MICCAI Challenge on Liver Ultrasound Tracking). It was awarded by the organizers of the CLUST challenges as being the best method for accurate target tracking in 3D ultrasound sequences. We recently improved our approach in order to increase its robustness to the presence of ultrasound shadows, local illumination changes and image occlusions.

7.6.2. 3D steering of flexible needle by ultrasound visual servoing

Participants: Pierre Chatelain, Jason Chevrie, Marie Babel, Alexandre Krupa.

The objective of this work is to provide robotic assistance during needle insertion procedures such as biopsy or ablation of localized tumor. In previous work, we designed a control approach based on a duty cycling technique for steering a beveled-tip flexible needle actuated by a robotic arm in such a way to control the needle curvature in 3D space and reach a desired target by visual servoing. In this preliminary work, the control approach was validated by using visual features extracted from 2 images provided by 2 orthogonal cameras observing a translucent gelatin phantom where the needle was inserted. This year, we have pursued our work towards this needle steering robotic assistance by developing a new algorithm able to track in real-time a flexible needle in a sequence of 3D ultrasound images (volumes). The flexible needle modeled as a polynomial curve is tracked during the automatic insertion using particle filtering. This new tracking algorithm enables real-time closed-loop needle control with 3D ultrasound feedback. The target to reach was manually defined by the user in the US image and can be on-line tracked thanks to the template tracking algorithm proposed in [21] based on ultrasound dense visual servoing [7]. Experimental results of an automatic needle tip positioning in a home-made gelatine phantom demonstrate the feasibility of 3D ultrasound-guided needle steering for reaching a desired target by ultrasound visual servoing [33]. Recently a new control law for needle steering that uses both direct manipulation of the needle base and the duty cycling method has been studied. It is based on a 3D model of a beveled tip needle using virtual springs that characterize the needle mechanical interaction with soft tissue. From this model, a measure of the controllability of the needle tip degrees of freedom was proposed in order to mix the control between the direct base manipulation and the duty cycling technique. Preliminary simulations show that this hybrid control allows better targeting capabilities in terms of larger needle workspace and reduced needle bending.

7.6.3. Optimization of ultrasound image quality by visual servoing

Participants: Pierre Chatelain, Alexandre Krupa.

This study focuses on a new ultrasound-based visual servoing approach that optimizes the positioning of an ultrasound probe manipulated by a robotic arm in order to improve the quality of the acquired ultrasound images. To this end, we use the recent framework of ultrasound confidence map, developed in the Chair for Computer Aided Medical Procedures and Augmented Reality of Prof. Nassir Navab, which aims at estimating the per-pixel quality of the ultrasound signal based on a model of sound propagation in soft tissues. More specifically, we treat the ultrasound confidence maps as a new modality and designed a visual servoing control law for image quality optimization. We illustrated our approach with the application of robotic tele-echography where the in-plane rotation of a 2D probe is visually servoed by the confidence map and the other degrees of

freedom are teleoperated by the user. Experiments performed on both an ultrasound examination training phantom and ex vivo tissue samples validated this new concept [32]. Currently, we consider the confidence-driven servoing of other degrees of freedom, in particular out-of-plane motions that were controlled in our previous works from image moments [6], which could provide finer control of the image quality.

7.6.4. Visual servoing based on ultrasound elastography

Participants: Pedro Alfonso Patlan Rosales, Alexandre Krupa.

This study concerns the use of the ultrasound elastography as a new image modality for the control of the motion of an ultrasound probe actuated by a robotic manipulator. Elastography imaging is performed by applying continuous stress variation on soft tissues in order to estimate a strain map of the observed tissues. It is obtained by estimating, from the RF (radio-frequency) signal along each scan line of the probe transducer, the echo time delays between pre- and post-compressed tissue. Usually, this continuous stress variation is performed manually by the user who manipulates the US probe and it results therefore in a user-dependent quality of the elastography image. To improve the US elastography imaging, we recently developed an assistant robotic palpation system that automatically moves an ultrasound probe in such a way to optimize ultrasound elastography. The main originality of this preliminary work concerns the use of the elastography modality directly as input of the robot controller thanks to an innovative ultrasound elastography-based visual servoing approach.

7.6.5. Visual servoing using shearlet transform

Participants: Lesley-Ann Duflot, Alexandre Krupa.

Similar to wavelet transform, shearlet transform is usually used in the field of signal or image compression. At the best of our knowledge these image representations were never used directly as feedback of a closed-loop control scheme. The objective of this work is to study the feasibility of using the coefficients of shearlet transform of the observed ultrasound image directly as the visual features of an image-based visual servoing. In this study we estimated numerically the interaction matrix that links the time variation of the coarsest coefficients of the shearlet to the motion of the ultrasound probe. This shearlet-based visual servoing was experimentally tested for automatically positioning a 2D US probe, held by a robot, on a desired section of an abdominal phantom. The first results demonstrated promising performances.

LINKMEDIA Project-Team

6. New Results

6.1. Unsupervised motif and knowledge discovery

6.1.1. Estimation of continuous intrinsic dimension

Participants: Laurent Amsaleg, Teddy Furon.

In collaboration with Michael Houle, National Institute for Informatics (Japan).

Some of our research work was concerned with the estimation of continuous intrinsic dimension (ID), a measure of intrinsic dimensionality recently proposed by Houle. Continuous ID can be regarded as an extension of Karger and Ruhl's expansion dimension to a statistical setting in which the distribution of distances to a query point is modeled in terms of a continuous random variable. This form of intrinsic dimensionality can be particularly useful in search, classification, outlier detection, and other contexts in machine learning, databases, and data mining, as it has been shown to be equivalent to a measure of the discriminative power of similarity functions. In [11], we proposed several estimators of continuous ID that we analyzed based on extreme value theory, using maximum likelihood estimation, the method of moments, probability weighted moments, and regularly varying functions. Experimental evaluation was performed using both real and artificial data.

6.1.2. Supervised multi-scale locality sensitive hashing

Participants: Laurent Amsaleg, Li Weng.

LSH is a popular framework to generate compact representations of multimedia data, which can be used for content based search. However, the performance of LSH is limited by its unsupervised nature and the underlying feature scale. In [42], we proposed to improve LSH by incorporating two elements: supervised hash bit selection and multi-scale feature representation. First, a feature vector is represented by multiple scales. At each scale, the feature vector is divided into segments. The size of a segment is decreased gradually to make the representation correspond to a coarse-to-fine view of the feature. Then each segment is hashed to generate more bits than the target hash length. Finally the best ones are selected from the hash bit pool according to the notion of bit reliability, which is estimated by bit-level hypothesis testing. Extensive experiments have been performed to validate the proposal in two applications: near-duplicate image detection and approximate feature distance estimation. We first demonstrate that the feature scale can influence performance, which is often a neglected factor. Then we show that the proposed supervision method is effective. In particular, the performance increases with the size of the hash bit pool. Finally, the two elements are put together. The integrated scheme exhibits further improved performance.

6.1.3. Rotation and translation covariant match kernels for image retrieval

Participants: Andrei Bursuc, Teddy Furon, Hervé Jégou, Giorgos Tolias.

Most image encodings achieve orientation invariance by aligning the patches to their dominant orientations and translation invariance by completely ignoring patch position or by max-pooling. Albeit successful, such choices introduce too much invariance because they do not guarantee that the patches are rotated or translated consistently. In this work, we propose a geometric-aware aggregation strategy, which jointly encodes the local descriptors together with their patch dominant angle [38] and/or location [10]. The geometric attributes are encoded in a continuous manner by leveraging explicit feature maps. Our technique is compatible with generic match kernel formulation and can be employed along with several popular encoding methods, in particular bag of words, VLAD and the Fisher vector. The method is further combined with an efficient monomial embedding to provide a codebook-free method aggregating local descriptors into a single vector representation. Invariance is achieved by efficient similarity estimation of multiple rotations or translations, offered by a simple trigonometric polynomial. This strategy is effective for image search, as shown by experiments performed on standard benchmarks for image and particular object retrieval, namely Holidays and Oxford buildings.

6.1.4. Sequential pattern mining on audio data

Participants: Laurent Amsaleg, Guillaume Gravier, Simon Malinowski.

M. Sc. Internship of Corentin Hardy, in collaboration with René Quiniou, Inria Rennes, DREAM research team, within the framework of the STIC AmSud Maximum project and of the MOTIF Inria Associate Team.

Analyzing multimedia data is a challenging problem due to the quantity and complexity of such data. Mining for frequently recurring patterns is a task often ran to help discovering the underlying structure hidden in the data. This year, we have explored how data symbolization and sequential pattern mining techniques could help for mining recurring patterns in multimedia data. In [20], we have shown that even if sequential pattern mining techniques are very helpful in terms of computational efficiency, the data symbolization step is a crucial step to find for extracting relevant audio patterns.

6.1.5. Clustering by diverting supervised machine learning

Participants: Vincent Claveau, Teddy Furon, Guillaume Gravier.

M. Sc. Internship of Amélie Royer, ENS Rennes.

Clustering algorithms exploit an input similarity measure on the samples, which should be fine-tuned with the data format and the application at hand. However, manually defining a suitable similarity measure is a difficult task in case of limited prior knowledge or complex data structures for example. While supervised classification systems require a set of samples annotated with their ground-truth classes, recent studies have shown it is possible to exploit classifiers trained on an artificial annotation of the data in order to induce a similarity measure. In this work, we have proposed a unified framework, named similarity by iterative classifications (SIC), which explores the idea of diverting supervised learning for automatic similarity inference. We studied several of its theoretical and practical aspects. We also have implemented and evaluate SIC on three tasks of knowledge discovery on multimedia content. Results show that in most situations the proposed approach indeed benefits from the underlying classifier's properties and outperforms usual similarity measures for clustering applications.

6.1.6. Multimodal person discovery in TV broadcasts

Participant: Guillaume Gravier.

Work in collaboration with Cassio Elias dos Santos Jr. and William Robson Schwartz, in the framework of the Inria Associate Team MOTIF and of the STIC AmSud project Maximum.

Taking advantage of recent results on large-scale face comparison with partial least square, we developed various approaches for multimodal person discovery in TV broadcasts in the framework of the MediaEval 2015 international benchmark [30]. The task consists in naming the persons on screen that are speaking with no prior information, leveraging text overlays, speech transcripts as well as face and voice comparison. We investigated two distinct aspects of multimodal person discovery. One refers to face clusters, which are considered to propagate names associated with faces in one shot to other faces that probably belong to the same person. The face clustering approach consists in calculating face similarities using partial least squares and a simple hierarchical approach. The other aspect refers to tag propagation in a graph-based approach where nodes are speaking faces and edges link similar faces/speakers. The advantage of the graph-based tag propagation is to not rely on face/speaker clustering, which we believe can be errorprone. The face clustering approach ranked among the top results in the international benchmark.

6.1.7. Unsupervised video structure mining with grammatical inference

Participants: Guillaume Gravier, Bingqing Qu.

In collaboration with Jean Carrive and Félicien Vallet, Institut National de l'Audiovisuel.

In [25], we addressed the problem of unsupervised program structuring with minimal prior knowledge about the program. We extended previous work to propose an approach able to identify multiple structures and infer structural grammars for recurrent TV programs of different types. The approach taken involves three sub-problems: i) we determine the structural elements contained in programs with minimal knowledge about which type of elements may be present; ii) we identify multiple structure for the programs if any and model the structures of programs; iii) we generate the structural grammar for each corresponding structure. Finally, we conducted use-case based evaluations on real recurrent programs of three different types to demonstrate the effectiveness of the proposed approach.

6.1.8. Information retrieval for distributional semantics, and vice-versa

Participants: Vincent Claveau, Ewa Kijak.

Distributional thesauri are useful in many tasks of natural language processing. In [33], [3], we address the problem of building and evaluating such thesauri with the help of information retrieval (IR) concepts. Two main contributions are proposed. First, in the continuation of previous work, we have shown how IR tools and concepts can be used with success to build thesauri. Through several experiments and by evaluating directly the results with reference lexicons, we show that some IR models outperform state-of-the-art systems. Secondly, we use IR as an application framework to indirectly evaluate the generated thesaurus. Here again, this task-based evaluation validate the IR approach used to build the thesaurus. Moreover, it allows us to compare these results with those from the direct evaluation framework used in the literature. The observed differences question these evaluation habits.

6.2. Multimedia content description and structuring

6.2.1. Image description using component trees

Participants: Petra Bosilj, Ewa Kijak.

In collaboration with Sébastien Lefèvre from Obelix Team (IRISA).

In this work, we explored the application of a tree-based feature extraction algorithm for the widely-used MSER features, and proposed a tree-of-shapes based detector of maximally stable regions. Changing an underlying component tree in the algorithm allows considering alternative properties and pixel orderings for extracting maximally stable regions. Performance evaluation was carried out on a standard benchmark in terms of repeatability and matching score under different image transformations, as well as in a large scale image retrieval setup, measuring mean average precision. The detector outperformed the baseline MSER in the retrieval experiments [37].

We also proposed a local region descriptor based on 2D shape-size pattern spectra, calculated on arbitrary connected regions, and combined with normalized central moments. The challenges when transitioning from global pattern spectra to the local ones were faced, and an exhaustive study on the parameters and the properties of the newly constructed descriptor was conducted. The descriptors were calculated on MSER regions, and evaluated in a simple retrieval system. Competitive performance with SIFT descriptors was achieved. An additional advantage of the proposed descriptors is their size which is less than half the size of SIFT [14], [15].

6.2.2. Improved motion description for action classification

Participant: Hervé Jégou.

In collaboration with Mihir Jain (University of Amsterdam, The Netherlands) and Patrick Bouthemy (Team-project SERPICO, Inria Rennes, France)

Even though the importance of explicitly integrating motion characteristics in video descriptions has been demonstrated by several recent papers on action classification, our current work concludes that adequately decomposing visual motion into dominant and residual motions, i.e., camera and scene motion, significantly improves action recognition algorithms. This holds true both for the extraction of the space-time trajectories and for computation of descriptors. We designed in [7] a new motion descriptor—the DCS descriptor—that captures additional information on local motion patterns enhancing results based on differential motion scalar quantities, divergence, curl and shear features. Finally, applying the recent VLAD coding technique proposed in image retrieval provides a substantial improvement for action recognition. These findings are complementary to each other and they outperformed all previously reported results by a significant margin on three challenging datasets: Hollywood 2, HMDB51 and Olympic Sports as reported in (Jain et al. (2013)).

6.2.3. Word embeddings and recurrent neural networks for spoken language understanding

Participants: Guillaume Gravier, Christian Raymond, Vedran Vukotić.

Recently, word embedding representations have been investigated for slot filling in spoken language understanding (SLU), along with the use of neural networks as classifiers. Neural networks, especially recurrent neural networks, which are adapted to sequence labeling problems, have been applied successfully on the popular ATIS database. In [29], we make a comparison of this kind of models with the previously state-of-the-art conditional random fields (CRF) classifier on a more challenging SLU database. We show that, despite efficient word representations used within these neural networks, their ability to process sequences is still significantly lower than for CRF, while also having a drawback of higher computational costs, and that the ability of CRF to model output label dependencies is crucial for SLU.

6.2.4. Hierarchical topic structuring

Participants: Guillaume Gravier, Pascale Sébillot, Anca-Roxana Şimon.

Topic segmentation traditionally relies on lexical cohesion measured through word re-occurrences to output a dense segmentation, either linear or hierarchical. We have proposed a novel organization of the topical structure of textual content [28]. Rather than searching for topic shifts to yield dense segmentation, our algorithm extracts topically focused fragments organized in a hierarchical manner. This is achieved by leveraging the temporal distribution of word re-occurrences, searching for bursts, to skirt the limits imposed by a global counting of lexical re-occurrences within segments. Comparison to a reference dense segmentation on varied datasets indicates that we can achieve a better topic focus while retrieving all of the important aspects of a text.

6.2.5. Partial least square hashing for large-scale face identification

Participants: Guillaume Gravier, Ewa Kijak.

Work performed with Cassio Elias dos Santos Jr. during his 3 months visit, in collaboration with William Robson Schwartz (UFMG, Brasil), in the framework of the Inria Associate Team MOTIF.

Face recognition has been largely studied in past years. However, most of the related work focus on increasing accuracy and/or speed to test a single pair probe-subject. In [31], we introduced a novel method inspired by the success of locality sensing hashing applied to large general purpose datasets and by the robustness provided by partial least squares analysis when applied to large sets of feature vectors for face recognition. The result is a robust hashing method compatible with feature combination for fast computation of a short list of candidates in a large gallery of subjects. We provided theoretical support and practical principles for the proposed hashing method that may be reused in further development of hash functions applied to face galleries. Comparative evaluations on the FERET and FRGCv1 datasets demonstrate a speedup of a factor 16 compared to scanning all subjects in the face gallery.

6.2.6. Selection strategies for active learning in NLP

Participants: Vincent Claveau, Ewa Kijak.

Nowadays, many NLP problems are modelized as supervised machine learning tasks, especially when it comes to information extraction. Consequently, the cost of the expertise needed to annotate the examples is a widespread issue. Active learning offers a framework to that issue, allowing to control the annotation cost while maximizing the classifier performance, but it relies on the key step of choosing which example will be proposed to the expert. In [3], we have examined and proposed such selection strategies in the specific case of conditional random fields which are largely used in NLP. On the one hand, we have proposed a simple method to correct a bias of certain state-of-the-art selection techniques. On the other hand, we have detailed an original approach to select the examples, based on the respect of proportions in the datasets. These contributions are validated over a large range of experiments implying several tasks and datasets, including named entity recognition, chunking, phonetization, word sense disambiguation.

6.2.7. Tree-structured named entities extraction from competing speech transcripts

Participant: Christian Raymond.

When real applications are working with automatic speech transcription, the first source of error does not originate from the incoherence in the analysis of the application but from the noise in the automatic transcriptions. In [41], we present a simple but effective method to generate a new transcription of better quality by combining utterances from competing transcriptions. We have extended a structured named entity (NE) recognizer submitted during the ETAPE challenge. Working on French TV and radio programs, our system revises the transcriptions provided by making use of the NEs it has detected. Our results suggest that combining the transcribed utterances which optimize the F-measure, rather than minimizing the WER scores, allows the generation of a better transcription for NE extraction. The results show a small but significant improvement of 0.9 % SER against the baseline system on the ROVER transcription. These are the best performances reported to date on this corpus.

6.3. Content-based information retrieval

6.3.1. A comparison of dense region detectors for image search and fine-grained classification

Participants: Hervé Jégou, Ahmet Iscen, Giorgos Tolias.

In collaboration with Philippe-Henri Gosselin (ETIS team, ENSEA, Cergy, France)

We consider a pipeline for image classification or search based on coding approaches like bag of words or Fisher vectors. In this context, the most common approach is to extract the image patches regularly in a dense manner on several scales. In [6], we propose and evaluate alternative choices to extract patches densely. Beyond simple strategies derived from regular interest region detectors, we propose approaches based on super-pixels, edges, and a bank of Zernike filters used as detectors. The different approaches are evaluated on recent image retrieval and fine-grain classification benchmarks. Our results show that the regular dense detector is outperformed by other methods in most situations, leading us to improve the state of the art in comparable setups on standard retrieval and fined-grain benchmarks. As a byproduct of our study, we show that existing methods for blob and super-pixel extraction achieve high accuracy if the patches are extracted along the edges and not around the detected regions.

6.3.2. Efficient large-scale similarity search using matrix factorization

Participants: Teddy Furon, Ahmet Iscen.

In collaboration with Michael Rabbat (McGill University, Montréal, Canada)

We considered the image retrieval problem of finding the images in a dataset that are most similar to a query image. Our goal is to reduce the number of vector operations and memory for performing a search without sacrificing accuracy of the returned images. We adopt a group testing formulation and design the decoding architecture using either dictionary learning or eigendecomposition. The latter is a plausible option for small-to-medium sized problems with high-dimensional global image descriptors, whereas dictionary learning is applicable in large-scale scenario. We evaluate our approach both for global descriptors obtained from SIFT and CNN features. Experiments with standard image search benchmarks, including the Yahoo100M dataset

comprising 100 million images, show that our method gives comparable (and sometimes superior) accuracy compared to exhaustive search while requiring only 10 % of the vector operations and memory. Moreover, for the same search complexity, our method gives significantly better accuracy compared to approaches based on dimensionality reduction or locality sensitive hashing [43].

6.3.3. Explicit embeddings for nearest neighbor search with Mercer kernels

Participant: Hervé Jégou.

In collaboration with Anthony Bourrier and Patrick Pérez (Technicolor, Rennes, France), Florent Perronnin (Xerox, Grenoble, France) Rémi Gribonval (Team-project PANAMA, Inria Rennes, France).

Many approximate nearest neighbor search algorithms operate under memory constraints, by computing short signatures for database vectors while roughly keeping the neighborhoods for the distance of interest. Encoding procedures designed for the Euclidean distance have attracted much attention in the last decade. In the case where the distance of interest is based on a Mercer kernel, we propose a simple, yet effective two-step encoding scheme: first, compute an explicit embedding to map the initial space into a Euclidean space; second, apply an encoding step designed to work with the Euclidean distance. Comparing this simple baseline with existing methods relying on implicit encoding, we demonstrate better search recall for similar code sizes with the chi-square kernel in databases comprised of visual descriptors, outperforming concurrent state-of-the-art techniques by a large margin [2].

6.3.4. Image search with selective match kernels: aggregation across single and multiple images

Participants: Hervé Jégou, Giorgos Tolias.

In collaboration with Yannis Avrithis (National Technical University of Athens, Greece)

Our work [9] considers a 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. 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 show that the same aggregation procedure, originally applied per image, can effectively operate on groups of similar features found across multiple images. This method implicitly performs feature set augmentation, while enjoying savings in memory requirements at the same time. Finally, the proposed method is shown effective for place recognition, outperforming state of the art methods on a large scale landmark recognition benchmark.

6.3.5. Early burst detection for memory-efficient image retrieval

Participant: Hervé Jégou.

In collaboration with Miajing Shi, visiting Ph. D. student from Pekin University, and Yannis Avrithis (National Technical University of Athens, Greece)

Recent works show that image comparison based on local descriptors is corrupted by visual bursts, which tend to dominate the image similarity. The existing strategies, like power-law normalization, improve the results by discounting the contribution of visual bursts to the image similarity. We proposed to explicitly detect the visual bursts in an image at an early stage. We compare several detection strategies jointly taking into account feature similarity and geometrical quantities. The bursty groups are merged into meta-features, which are used as input to state-of-the-art image search systems such as VLAD or the selective match kernel. Then, we show the interest of using this strategy in an asymmetrical manner, with only the database features being aggregated but not those of the query. Extensive experiments performed on public benchmarks for visual retrieval show the benefits of our method, which achieves performance on par with the state of the art but with a significantly reduced complexity, thanks to the lower number of features fed to the indexing system [40], [44].

6.3.6. Biomedical information retrieval

Participants: Vincent Claveau, Ewa Kijak.

In collaboration with N. Grabar (STL), T. Hamon (LIMSI), and S. Le Maguer (Univ. Saarland).

The right of patients to access their clinical health record is granted by the code of Santé Publique. Yet, this piece of content remains difficult to understand. We propose different IR experiments in which we use queries defined by patients in order to find relevant documents [3], [16]. We use the Indri search engine, based on statistical language modeling, as well as semantic resources. More precisely, our approaches are chiefly based on the terminological variation (e.g., synonyms, abbreviations) to link between expert and patient languages. Various combinations of resources and Indri settings are explored, mostly based on query expansion.

6.4. Linking, navigation and analytics

6.4.1. Sentiment analysis on social networks

Participants: Vincent Claveau, Christian Raymond, Vedran Vukotić.

In the framework of our participation to the DeFT 2015 text-mining challenge, we have developped sentiment-analysis methods for tweets [34]. Several sub-tasks have been considered: i) valence classification of tweets and ii) fine-grained classification of tweets (which includes two sub-tasks: detection of the generic class of the information expressed in a tweet and detection of the specific class of the opinion/sentiment/emotion. For all three problems, we adopt a standard machine learning framework. More precisely, three main methods are proposed and their feasibility for the tasks is analyzed: i) decision trees with boosting (bonzaiboost), ii) naive Bayes with Okapi and iii) convolutional neural networks (CNNs). Our approaches are voluntarily knowledge free and text-based only, we do not exploit external resources (lexicons, corpora) or tweet metadata. It allows us to evaluate the interest of each method and of traditional bag-of-words representations vs. word embeddings. Methods using simple ML frameworks and IR-based similarity metrics have been demonstrated to yield the best results.

6.4.2. A multi-dimensional data model for personal photo browsing

Participant: Laurent Amsaleg.

Work performed in the framework of the CNRS PICS MMAnalytics, and in collaboration with Marcel Worring, Univeristy of Amsterdam (The Netherlands)

Digital photo collections—personal, professional, or social—have been growing ever larger, leaving users overwhelmed. It is therefore increasingly important to provide effective browsing tools for photo collections. Learning from the resounding success of multi-dimensional analysis (MDA) in the business intelligence community for on-line analytical processing (OLAP) applications, we proposed a multi-dimensional model for media browsing, called M3, that combines MDA concepts with concepts from faceted browsing [21]. We present the data model and describe preliminary evaluations, made using server and client prototypes, which indicate that users find the model useful and easy to use.

6.4.3. NLP-driven hyperlink construction in broadcast videos

Participants: Rémi Bois, Guillaume Gravier, Pascale Sébillot, Anca-Roxana Şimon.

In collaboration with Sien Moens (Katholieke Universiteit Leuven, Belgium), Éric Jamet and Martin Ragot (Univ. Rennes 2, France).

In the context of the the CominLabs project "Linking media in acceptable hypergraphs" dedicated to the creation of explicit and meaningful links between multimedia documents or fragments of documents, we have introduced a typology of possible links between contents of a multimedia news corpus [32]. While several typologies have been proposed and used by the community, we argue that they are not adapted to rich and large corpora which can contain texts, videos, or radio stations recordings. We have defined a new typology, as a first step towards automatically creating and categorizing links between documents' fragments in order to create new ways to navigate, explore, and extract knowledge from large collections.

We also investigated video hyperlinking based on speech transcripts, leveraging a hierarchical topical structure to address two essential aspects of hyperlinking, namely, serendipity control and link justification [26]. We proposed and compared different approaches exploiting a hierarchy of topic models as an intermediate representation to compare the transcripts of video segments. These hierarchical representations offer a basis to characterize the hyperlinks, thanks to the knowledge of the topics which contributed to the creation of the links, and to control serendipity by choosing to give more weights to either general or specific topics. Experiments have been performed on BBC videos from the Search and Hyperlinking task at MediaEval. Link precisions similar to those of direct text comparison have been achieved however exhibiting different targets along with a potential control of serendipity.

The Search and Anchoring in Video Archives task at MediaEval addressed two issues: The Search part aims at returning a ranked list of video segments that are relevant to a textual user query; The Anchoring part focuses on identifying video segments that would encourage further exploration within the archive. Capitalizing on the experience acquired in previous participations, we implemented a two step approach for both sub-tasks [27]. The first step, common to both, consists in generating a list of potential anchor segments and response-query segments relying on a hierarchical topical structuring technique. In the second step, for each query, the best 20 segments are selected according to content-based comparisons, while for the anchor detection sub-task, the segments are ranked based on a cohesion measure. The use of a hierarchical topical structure helps to propose segments of variable length at different levels of details with precise jump-in points for them. More, the algorithm deriving the structure relies on the burstiness phenomenon in word occurrences which gives an advantage over the classical bag-of-words model.

6.4.4. Information extraction

Participants: Vincent Claveau, Ewa Kijak.

In collaboration with X. Tannier (LIMSI), A. Vilnat (LIMSI) and B. Arnulphy (ANR).

Identifying events from texts is an information extraction task necessary for many NLP applications. Through the TimeML specifications and TempEval challenges, it has received some attention in the last years; yet, no reference result is available for French. In [12], we try to fill this gap by proposing several event extraction systems, combining for instance Conditional Random Fields, language modeling and k-nearest-neighbors. These systems are evaluated on French corpora and compared with state-of-the-art methods on English. The very good results obtained on both languages validate our whole approach.

6.5. Participation in benchmarking initiatives

- Video hyperlinking, TRECVid
- Search and anchoring, Mediaeval Multimedia International Benchmark
- Multimodal person discovery in broadcast TV, Mediaeval Multimedia International Benchmark
- DeFT 2015 text-mining challenge

MIMETIC Project-Team

7. New Results

7.1. Biomechanics for motion analysis-synthesis

Participants: Charles Pontonnier, Georges Dumont, Steve Tonneau, Franck Multon, Julien Pettré, Ana Lucia Cruz Ruiz, Antoine Muller.

Ana-Lucia Cruz-Ruiz has been recruited as a PhD student since november 2013. The goal of this thesis is to define and evaluate muscle-based controllers for avatar animation. We developed an original control approach to reduce the redudancy of the musculoskeletal system for motion synthesis, based on the muscle synergy theory. For this purpose we ran an experimental campaign of overhead throwing motions. We recorded the muscle activity of 10 muscles of the arm and the motion of the subjects. Thanks to a synergy extraction algorithm, we extracted a reduced set of activation signals corresponding to the so called muscle synergies and used them as an input in a forward dynamics pipeline . Thanks to a two stage optimization method, we adapted the model's muscle parameters and the synergy signals to be as close as possible of the recorded motion. The results are compelling and ask for further developments [9], [24].

We are also developing an analysis pipeline thanks to the work of Antoine Muller. This pipeline aims at using a modular and multiscale description of the human body to let users be able to analyse human motion. For now, the pipeline is able to assemble different biomechanical models in a convenient descriptive graph [15], Calibrate those models thanks to experimental data [30] and run inverse dynamics to get joint torques from exeperimental motion capture data [14].

7.2. VR and Ergonomics

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

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.

We deeply assessed the propensity of a virtual reality immersive room and classical interaction devices to evaluate properly the physical risk factors associated to assembly tasks. For this purpose, we compared tasks realized in real and virtual environment in terms of shoulder kinematics and muscular activity [20] and in terms of controlled kinematical variables, on the basis of the uncontrolled manifold theory [31]. Results show that there is less difference between real and virtual conditions than between indviduals, that make us think that such a virtual environment can be used to assess this type of task.

7.3. Interactions between walkers

Participants: Anne-Hélène Olivier, Armel Crétual, Julien Bruneau, Richard Kulpa, Sean Lynch, Julien Pettré.

Interaction between people, and especially local interaction between walkers, is a main research topic of MimeTIC. We propose experimental approaches using both real and virtual environments. This year, we developped new experiments in our immersive platform. First, we investigated obstacle avoidance behavior during real walking in a large immersive projection setup [22]. We analyze the walking behavior of users when avoiding real and virtual static obstacles. Indeed, CAVE-like immersive projection environments enable users to see both virtual and real objects, including the user's own body. With recent advances in VR technologies it becomes possible to build large-scale tracked immersive projection environments, which enable users to control their position in a large region of interest by real walking. In such environments virtual and real objects as well as multiple users or avatars may coexist in the same interaction space. Hence, it becomes important to gain an understanding of how the user's behavior is affected by the differences in perception and affordances of such real and virtual obstacles. We consider both anthropomorphic and inanimate objects,

each having his virtual and real counterpart. The results showed that users exhibit different locomotion behaviors in the presence of real and virtual obstacles, and in the presence of anthropomorphic and inanimate objects. Precisely, the results showed a decrease of walking speed as well as an increase of the clearance distance (i. e., the minimal distance between the walker and the obstacle) when facing virtual obstacles compared to real ones. Moreover, users act differently due to their perception of the obstacle: users keep more distance when the obstacle is anthropomorphic compared to an inanimate object and when the orientation of anthropomorphic obstacle is from the profile compared to a front position. However, although we observed differences in collision avoidance behavior between real and virtual obstacles, which indicate biases of natural locomotion introduced by the setup, their magnitude seem lower compared to typical results found in HMD environments.This suggests that although the user's behavior in mixed environments varies depending on the nature of the stimulus, the user's locomotion behavior and the management of his/her interaction space is comparable with the ones in real life. Considering these findings, our results open promising vistas for using large CAVE-like setups for socio-physical experiments, in particular in the fields of locomotion and behavioral dynamics.

Second, we studied interactions between an individual and a crowd [7]. When avoiding a group, a walker has two possibilities: either he goes through it or around it. Going through very dense group or around huge one would not seem natural and could break any sense of presence in a virtual environment. The aim of this work was to enable crowd simulators to correctly handle such situations. To this end, we need understanding how real humans decide to go through or around groups. As a first hypothesis, we apply the Principle of Minimum Energy (PME) on different group sizes and density. According to it, a walker should go around small and dense groups while he should go through large and sparse groups. We quantified decision thresholds. However, PME left some inconclusive situations for which the two solutions paths have similar energetic cost. In a second part, we proposed an experiment to corroborate PME decisions thresholds with real observations. We proposed using Virtual Reality to enable accurately controlling experimental factors. We considered as well the role of secondary factors in inconclusive situations. We showed the influence of the group appearance and direction of relative motion in the decision process. Finally, we draw some guidelines to integrate our conclusions to existing crowd simulators and demonstrate that spectators can perceive some improvement in the crowd animation.

This year, we also developed new experiments in real conditions by considering the interaction between a walker and a moving robot. This worked was performed in collaboration with Philippe Souères and Christian Vassallo (LAAS, Toulouse). The development of Robotics accelerated these recent years, it is clear that robots and humans will share the same environment in a near future. In this context, understanding local interactions between humans and robots during locomotion tasks is important to steer robots among humans in a safe manner. Our work is a first step in this direction. Our goal is to describe how, during locomotion, humans avoid collision with a moving robot. We study collision avoidance between participants and a non-reactive robot (we wanted to avoid the effect of a complex loop by a robot reacting to participants' motion). Our objective is to determine whether the main characteristics of such interaction preserve the ones previously observed: accurate estimation of collision risk, anticipated and efficient adaptations. We observed that collision avoidance between a human and a robot has similarities with human-human interactions (estimation of collision risk, anticipation) but also leads to major differences. Humans preferentially give way to the robot, even if this choice is not optimal with regard to motion adaptation to avoid the collision. We proposed to interpret this behavior based on the notion of perceived danger and safety. Given the difficulty to understand how a robot behaves, and the lack of experience of interactions with the robot, humans apply a conservative avoidance strategy and prefer giving way to the robot. However, it is important to note that human participants succeed in perceiving the motion of the robot (anticipation was observed, no aberrant reaction occurred). One main conclusion is that, if we control robots to move like humans, we have a risk facing unexpected situations where robot compensates and cancels humans adaptations to the robot. A robot programmed to be cooperative could be perceived as hostile. The conclusion of this study opens paths for future research. A first direction is to better understand the possible effect of this notion of danger during interactions. We believe that this notion is of even higher importance when studying interactions with vehicles: a risk of collision with a fast vehicle obviously raises higher danger. A second direction is about the design of safe robots moving among human walkers. How

the robot should adapt to others? Should it be collaborative with the risk of compensating human avoidance strategies? Should it be passive? We believe that robots should first be equipped with the ability to early detect humans avoidance strategy and adapt to it. In the near future, we want to continue our study of interactions between a robot and a human. In a first step, we plan to equip the robot with collision avoidance system which imitates real human strategies, and investigate how participants adapt to this new situation in comparison with a passive robot.

Finally, Sean Dean Lynch has been recruited as a PhD student since september 2015. This thesis concerns the visual perception of human motion during interactions in locmotor tasks. From the visual perception of someone's motion, we are able to predict the future course of this motion, interpret and anticipate his/her intentions and adapt our own motion to allow interactions. The main objective of the thesis is to identify the underlying perceptual mechanisms, i.e., the human motion cues which are necessary for an accurate understanding of others' intentions. It would allow to make significant progress in the understanding of human social behaviors. To reach these objectives, the thesis will be based on an experimental approach in virtual reality.

7.4. Motion Sensing

Participants: Franck Multon, Pierre Plantard.

Recording human activity is a key point of many applications and fundamental works. Numerous sensors and systems have been proposed to measure positions, angles or accelerations of the user's body parts. Whatever the system is, one of the main is to be able to automatically recognize and analyze the user's performance according to poor and noisy signals. Hence, recognizing and measuring human performance are important scientific challenges especially when using low-cost and noisy motion capture systems. MimeTIC has addressed the above problems in two main application domains.

Firstly, in ergonomics, we explored the use of low-cost mottion capture systems, a Microsoft Kinect, to measure the 3D pose of a subject in natural environments, such as on a workstation, with many occlusions and inappropriate sensor placements. Predicting the potential accuracy of the measurement for such complex 3D poses and sensor placements is challenging with classical experimental setups. To tackle this problem, we propose [16] a new evaluation method based on a virtual mannequin. Thanks to this evaluation method, more than 500,000 configurations have been automatically tested, which is almost impossible to evaluate with classical protocols. The results show that the kinematic information obtained by the Kinect system is generally accurate enough to fill-in ergonomic assessment grids. However inaccuracy strongly increases for some specific poses and sensor positions. Using this evaluation method enabled us to report configurations that could lead to these high inaccuracies. Results obtained with the virtual mannequin are in accordance with those obtained with a real subject for a limited set of poses and sensor configuration. This knowledge can help to anticipate potential problems using a Kinect in given scenarios, and to propose methods to tackle these expected problems.

Secondly, in clinical gait analysis, we proposed a method to overcome the main limitations imposed by the low accuracy of the Kinect measurements in real medical exams. Indeed, inaccuracies in the 3D depth images leads to badly reconstructed poses and inaccurate gait event detection. In the latter case, confusion between the foot and the ground leads to inaccuracies in the foot-strike and toe-off event detection, which are essential information to get in a clinical exam. To tackle this problem we assumed that heel strike events could be indirectly estimated by searching for the extreme values of the distance between the knee joints along the walking longitudinal axis [5] . As Kinect sensor may not accurately locate the knee joint, we used anthropometrical data to select a body point located at a constant height where the knee should be in the reference posture. Compared to previous works using a Kinect, heel strike events and gait cycles are more accurately estimated, which could improve global clinical gait analysis frameworks with such a sensor. Once these events are correctly detected, it is possible to define indexes that enables the clinician to have a rapid state of the quality of the gait. We proposed [4] a new method to asses gait asymmetry based on depth images, to decrease the impact of errors in the Kinect joint tracking system. It is based on the longitudinal

spatial difference between lower-limb movements during the gait cycle. The movement of artifically impaired gaits was recorded using both a Kinect placed in front of the subject and a motion capture system. The proposed longitudinal index distinguished asymmetrical gait ($p < 0.001$), while other symmetry indices based on spatiotemporal gait parameters failed using such Kinect skeleton measurements. This gait asymmetry index measured with a Kinect is low cost, easy to use and is a promising development for clinical gait analysis.

7.5. Virtual Human Animation

Participants: Julien Pettré, Franck Multon, Steve Tonneau.

Multiped locomotion in cluttered environments is addressed as the problem of planning acyclic sequences of contacts, that characterize the motion. In order

to overcome the inherent combinatorial difficulty of the problem, we separate it in two subproblems [34]: first, planning a guide trajectory for the root of the robot and then, generating relevant contacts along this trajectory. This paper proposes theoretical contributions to these two subproblems. We propose a theoretical characterization of the guide trajectory, named "true feasibility", which guarantee that a guide can be mapped into the contact manifold of the robot. As opposed to previous approaches, this property makes it possible to assert the relevance of a guide trajectory without explicitly computing contact configurations, as proposed in our previous works. This property can be efficiently checked by a sample-based planner (e.g. we implemented a visibility PRM). Since the guide trajectories that we characterized are easily mapped to a valid sequence of contacts, we then focused on how to select a particular sequence with desirable properties, such as robustness, efficiency and naturalness, only considered for cyclic locomotion so far. Based on these novel theoretical developments, we implemented a complete acyclic contact planner and demonstrate its efficiency by producing a large variety of movements with three very different robots (humanoid, insectoid, dexterous hand) in five challenging scenarios. The planner is very efficient in quality of the produced movements and in computation time: given a computed RB-PRM, a legged figure or a dexterous hand can generate its motion in real time. This result outperforms any previous acyclic contact planner.

7.6. VR and sports

Participants: Richard Kulpa, Benoit Bideau, Franck Multon, Anne-Hélène Olivier.

Athletes' performances are influenced by internal and external factors, including their psychological state and environmental factors, especially during competition. As a consequence, current training programs include stress management. In this work [3], we explore whether highly immersive systems can be used for such training programs. First, we propose methodological guidelines to design sport training scenarios both on considering the elements that a training routine must have and how external factors might influence the participant. The proposed guidelines are based on Flow and social-evaluative threat theories. Second, to illustrate and validate our methodology, we designed an experimental setup reproducing a 10 m Olympic pistol shooting. We analyzed whether changes in the environment are able to induce changes in user performance, physiological responses, and the subjective perception of the task. The simulation included stressors in order to raise a social-evaluative threat, such as aggressive public behavior or unforced errors, increasing the pressure while performing the task. The results showed significant differences in their subjective impressions, trends in the behavioral and physiological data were also observed. Taken together, our results suggest that highly immersive systems could be further used for training in sports.

Among the stimuli, visual information uptake is a fundamental element of sports involving interceptive tasks. Several methodologies, like video and methods based on virtual environments, are currently employed to analyze visual perception during sport situations. Both techniques have advantages and drawbacks. We made an experiment to determine which of these technologies may be preferentially used to analyze visual information uptake during a sport situation [21]. To this aim, we compared a handball goalkeeper's performance using two standardized methodologies: video clip and virtual environment. We examined this performance for two response tasks: an uncoupled task (goalkeepers show where the ball ends) and a coupled task (goalkeepers

try to intercept the virtual ball). Variables investigated in this study were percentage of correct zones, percentage of correct responses, radial error and response time. The results showed that handball goalkeepers were more effective, more accurate and started to intercept earlier when facing a virtual handball thrower than when facing the video clip. These findings suggested that the analysis of visual information uptake for handball goalkeepers was better performed by using a 'virtual reality'-based methodology.

In a previous work, we analyzed the performance of beginners as they shot basketball free throws using various immersive conditions. Our results supported the assumption that natural complex motor behavior is possible in a VE, with little motor adaptation. The ultimate goal of our work is to design a VE training system for basketball free throws, so in this article we compare the performance of beginners making free throws in various visual conditions (first- versus third-person views using a large-screen immersive display) with that of expert players in the real world [8]. The key idea is to analyze how different visual conditions affect the performance of novices and to what extent it enables them to match the experts' performance.

Distance underestimation or any other perceptual disturbance in VR makes people adapt to the task at hand. The users in our study reached the same success rate by finding a new way to throw the ball, despite this incongruity between perception and action. The main observations reported in this article reinforce the conclusions in previous work, stating that 3PP is more efficient for certain tasks, but further work is required to test this result against other types of training conditions.

Finally, we worked on a transportable virtual reality system to analyse sports situations [6]. We proposed an original methodology to study the action of a goalkeeper facing a free kick. This methodology is based on a virtual reality setup in which a real goalkeeper is facing a virtual player and a virtual defensive wall. The setup has been improved to provide a total freedom of movement to the goalkeeper in order to have a realistic interaction between the goalkeeper and the player. The goalkeeper's movements are captured in real-time to accurately analyze his reactions. Such a methodology not only represents a valuable research tool but also provides a relevant training tool. Using this setup, this paper shows that goalkeepers are more performant during free kick with a wall composed of 5 defenders whatever its position.

7.7. Scheduling activities under spatial and temporal constraints

Participants: Fabrice Lamarche, Carl-Johan Jorgensen.

This work focusses on generating statistically consistent behaviors that can be used to pilot crowd simulation models over long periods of time, up to multiple days [1]. In real crowds, people's behaviors mainly depend on the activities they intend to perform. The way this activity is scheduled rely on the close interaction between the environment, space and time constraints associated with the activity and personal characteristics of individuals. Compared to the state of the art, our model better handle this interaction.

Our main contributions lie in the cdomain of activity scheduling and path planning. First, we proposed an individual activity scheduling process and its extension to cooperative activity scheduling. Based on descriptions of the environment, of intended activities and of agents' characteristics, these processes generate a task schedule for each agent. Locations where the tasks should be performed are selected and a relaxed agenda is produced. This task schedule is compatible with spatial and temporal constraints associated with the environment and with the intended activity of the agent and of other cooperating agents. It also takes into account the agents personal characteristics, inducing diversity in produced schedules. We showed that this model produces schedules statistically coherent with the ones produced by humans in the same situations. Second, we proposed a hierarchical path-planning process. It relies on an automatic environment analysis process that produces a semantically coherent hierarchical representation of virtual cities. The hierarchical nature of this representation is used to model different levels of decision making related to path planning. A coarse path is first computed, then refined during navigation when relevant information is available. It enable the agent to seamlessly adapt its path to unexpected events. Finally, those models have been included in a simulation platform that is able to simulate several thousand of pedestrians performing their daily activities in real-time. In order to deal with unexpected events, a process enabling adaptations of the pedestrian behavior have been designed. Those adaptations range from path modification to schedule adaptation according to the observed situation.

The proposed model handles long term rational decisions driving the navigation of agents in virtual cities. It considers the strong relationship between time, space and activity to produce more credible agents' behaviors. It can be used to easily populate virtual cities in which observable crowd phenomena emerge from individual activities.

7.8. Shoulder biomechanics

Participant: Armel Crétual [contact].

Shoulder hyperlaxity (SHL) is considered a main risk factor for shoulder instability and can be associated with different clinical shoulder instability presentations, such a multidirectional instability or unstable painful shoulder. Interestingly, quantification of shoulder laxity and hyperlaxity, particularly during physical examination, still remains an unsolved problem. Indeed, it is still frequently evaluated only through mono-axial amplitude, in particular using external rotation of the arm whilst at the side (ER1). We previously showed that this parameter is sensitive to inter-operator variability.

Therefore, we proposed a novel way to account for global shoulder mobility, the Shoulder Configuration Space Volume (SCSV) corresponding to the reachable volume in the configuration space of the shoulder joint [10]. In mechanics and robotics, the configuration space is the set of all reachable combination of coordinates. Considering the shoulder as the single joint between thorax and humerus instead of a combination of 4 actual joints (gleno-humeral, thoraco-humeral, scapulo-thoracic and sterno-clavicular), these coordinates are based upon the three joint angles defined by the International Society of Biomechanics (ISB) recommendations as plane of elevation orientation, elevation and axial rotation.

Then, this new index was examined through correlation to shoulder signs of hyperlaxity [19] for which we have shown a link with instability in patients who received a surgical procedure [18].

7.9. The Toric Space: a novel representation for camera control applications

Participants: Marc Christie, Christophe Lino, Quentin Galvane.

Many types of computer graphics applications such as data visualization or virtual movie production require users to position and move viewpoints in 3D scenes to effectively convey visual information or tell stories. The desired viewpoints and camera paths need to satisfy a number of visual properties (e.g. size, vantage angle, visibility, and on-screen position of targets). Yet, existing camera manipulation tools only provide limited interaction methods and automated techniques remain computationally expensive.

We introduce the *Toric space*, a novel and compact representation for intuitive and efficient virtual camera control. We first show how visual properties are expressed in this Toric space and propose an efficient interval-based search technique for automated viewpoint computation. We then derive a novel screen-space manipulation technique that provides intuitive and real-time control of visual properties. Finally, we propose an effective viewpoint interpolation technique which ensures the continuity of visual properties along the generated paths. The proposed approach (i) performs better than existing automated viewpoint computation techniques in terms of speed and precision, (ii) provides a screen-space manipulation tool that is more efficient than classical manipulators and easier to use for beginners, and (iii) enables the creation of complex camera motions such as long takes in a very short time and in a controllable way. As a result, the approach should quickly find its place in a number of applications that require interactive or automated camera control such as 3D modelers, navigation tools or games. The paper has been presented at SIGGRAPH 2015 (see [12] for more details).

We then rely on this Toric Space representation to contruct optimal camera paths (optimal in the satisfaction of visual properties along the path). Indeed, when creating real or computer graphics movies, the questions of how to layout elements on the screen, together with how to move the cameras in the scene are crucial to properly conveying the events composing a narrative. Though there is a range of techniques to automatically compute camera paths in virtual environments, none have seriously considered the problem of generating realistic camera motions even for simple scenes. Among possible cinematographic devices, real cinematographers often rely on camera rails to create smooth camera motions which viewers are familiar with. Following

this practice, we have proposed a method for generating virtual camera rails and computing smooth camera motions on these rails. Our technique analyzes characters motion and user-defined framing properties to compute rough camera motions which are further refined using constrained-optimization techniques. Comparisons with recent techniques demonstrate the benefits of our approach and opens interesting perspectives in terms of creative support tools for animators and cinematographers. See [25] for more details.

TO address the more general problem of solving contradicting visual properties, novel ways of aggregating functions has also been proposed [33].

7.10. Data-driven Virtual Cinematography

Participant: Marc Christie.

Our propelling motivation here is to rely on existing data from real movies (automatically extracted or manually annotated), to propose better better and better framing techniques.

We first contributed to the problem of automated editing, by reproducing elements of cinematographic style. Automatically computing a cinematographic consistent sequence of shots over a set of actions occurring in a 3D world is a complex task which requires not only the computation of appropriate shots (viewpoints) and appropriate transitions between shots (cuts), but the ability to encode and reproduce elements of cinematographic style. Models proposed in the literature, generally based on finite state machine or idiom-based representations, provide limited functionalities to build sequences of shots. These approaches are not designed in mind to easily learn elements of cinematographic style, nor do they allow to perform significant variations in style over the same sequence of actions. We have proposed a model for automated cinematography that can compute significant variations in terms of cinematographic style, with the ability to control the duration of shots and the possibility to add specific constraints to the desired sequence. The model is parameterized in a way that facilitates the application of learning techniques. By using a Hidden Markov Model representation of the editing process, we have demonstrated the possibility of easily reproducing elements of style extracted from real movies. Results comparing our model with state-of-the-art first order Markovian representations illustrate these features, and robustness of the learning technique is demonstrated through cross-validation. See [13] for more details.

We also proposed a tool to ease the process of annotating cinematographic content, for the purposes of both film analysis, and film synthesis [29]. The work relies on the proposition of a film language that extends previous representations such as PSL (Prose Storyboard Language) by integrating the editing aspects, through the notion of cinematographic “techniques” described as patterns of shots.

The proposed language, named “Patterns”, is described in more details in [35]. Our language can express the aesthetic properties of framing and shot sequencing, and of camera techniques used by real directors. Patterns can be seen as the semantics of camera transitions from one frame to another. The language takes an editors view of on-screen aesthetic properties: the size, orientation, relative position, and movement of actors and objects across a number of shots. We have illustrated this language through a number of examples and demonstrations. Combined with camera placement algorithms, we demonstrated the language’s capacity to create complex shot sequences in data-driven generative systems for 3D storytelling applications.

7.11. Logic control in interactive storytelling

Participants: Marc Christie, Hui-Yin Wu.

With the rising popularity of engaging storytelling experiences in gaming arises the challenge of designing logic control mechanisms that can adapt to increasingly interactive, immersive, and dynamic 3D gaming environments. Currently, branching story structures are a popular choice for game narratives, but can be rigid, and authoring mistakes may result in dead ends at runtime. This calls for automated tools and algorithms for logic control over flexible story graph structures that can check and maintain authoring logic at a reduced cost while managing user interactions at runtime. In this work we introduce a graph traversal method for logic control over branching story structures which allow embedded plot lines. The mechanisms are designed to

assist the author in specifying global authorial goals, evaluating the sequence of events, and automatically managing story logic during runtime. Furthermore, we showed how our method can be easily linked to 3D interactive game environments through a simple example involving a detective story with a flashback. See [36] for more details.

7.12. Automatic Continuity Editing for 3-D Animation

Participants: Marc Christie, Quentin Galvane, Christophe Lino.

We have proposed an optimization-based approach for automatically creating movies from 3-D animation. The method nicely separates the work of the virtual cinematographer (placing cameras and lights to produce nice-looking views of the action) from the work of the virtual film editor (cutting and pasting shots from all available cameras). While previous work has mostly focused on the first problem, the second problem has never been addressed in full details. We have reviewed the main causes of editing errors and built a cost function for minimizing them. We made a plausible semi-Markov assumption, which results in a computationally efficient dynamic programming solution. We showed that our method generates movies that avoid many common errors in film editing, including jump cuts, continuity errors and non-motivated cuts. We also show that our method can generate movies with different pacings. Combined with state-of-the-art cinematography, our approach therefore promises to significantly extend the expressiveness and naturalness of virtual movie-making. The work has been published at AAAI [27]. More details comparisons have been performed in [26].

PANAMA Project-Team

7. New Results

7.1. 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 9.2.1.1).

7.1.1. Theoretical results on sparse representations, graph signal processing, and dimension reduction

Participants: Rémi Gribonval, Yann Traonmilin, Gilles Puy, Nicolas Tremblay, Pierre Vandergheynst.

Main collaboration: Mike Davies (University of Edinburgh), Pierre Borgnat (ENS Lyon),

Stable recovery of low-dimensional cones in Hilbert spaces: Many inverse problems in signal processing deal with the robust estimation of unknown data from underdetermined linear observations. Low dimensional models, when combined with appropriate regularizers, have been shown to be efficient at performing this task. Sparse models with the ℓ_1-norm or low rank models with the nuclear norm are examples of such successful combinations. Stable recovery guarantees in these settings have been established using a common tool adapted to each case: the notion of restricted isometry property (RIP). This year, we established generic RIP-based guarantees for the stable recovery of cones (positively homogeneous model sets) with arbitrary regularizers. These guarantees were illustrated on selected examples. For block structured sparsity in the infinite dimensional setting, we used the guarantees for a family of regularizers which efficiency in terms of RIP constant can be controlled, leading to stronger and sharper guarantees than the state of the art. A journal paper is currently under revision [57].

Recipes for stable linear embeddings from Hilbert spaces to $\mathbb{R}^m$: We considered the problem of constructing a linear map from a Hilbert space (possibly infinite dimensional) to $\mathbb{R}^m$ that satisfies a restricted isometry property (RIP) on an arbitrary signal model set. We obtained a generic framework that handles a large class of low-dimensional subsets but also *unstructured* and *structured* linear maps. We provided a simple recipe to prove that a random linear map satisfies a general RIP on the model set with high probability. We also described a generic technique to construct linear maps that satisfy the RIP. Finally, we detailed how to use our results in several examples, which allow us to recover and extend many known compressive sampling results. This has been presented at the conference EUSIPCO 2015 [28], and a journal paper has been submitted [55].

Random sampling of bandlimited signals on graphs: We studied the problem of sampling k-bandlimited signals on graphs. We proposed two sampling strategies that consist in selecting a small subset of nodes at random. The first strategy is non-adaptive, i.e., independent of the graph structure, and its performance depends on a parameter called the graph coherence. On the contrary, the second strategy is adaptive but yields optimal results. Indeed, no more than $O(klog(k))$ measurements are sufficient to ensure an accurate and stable recovery of all k-bandlimited signals. This second strategy is based on a careful choice of the sampling distribution, which can be estimated quickly. Then, we proposed a computationally efficient decoder to reconstruct k-bandlimited signals from their samples. We proved that it yields accurate reconstructions and that it is also stable to noise. Finally, we conducted several experiments to test these techniques. A journal paper has been submitted [56].

Accelerated spectral clustering: We leveraged the proposed random sampling technique to propose a faster spectral clustering algorithm. Indeed, classical spectral clustering is based on the computation of the first k eigenvectors of the similarity matrix' Laplacian, whose computation cost, even for sparse matrices, becomes prohibitive for large datasets. We showed that we can estimate the spectral clustering distance matrix without computing these eigenvectors: by graph filtering random signals. Also, we took advantage of the stochasticity of these random vectors to estimate the number of clusters k. We compared our method to classical spectral clustering on synthetic data, and show that it reaches equal performance while being faster by a factor at least two for large datasets. A conference paper has been accepted at ICASSP 2016 [43] and a long version is in preparation.

7.1.2. Algorithmic and theoretical results on dictionary learning

Participants: Rémi Gribonval, Luc Le Magoarou, Nicolas Bellot, Thomas Gautrais, Nancy Bertin, Srdan Kitic.

Main collaboration (theory for dictionary learning): Rodolphe Jenatton, Francis Bach (Equipe-projet SIERRA (Inria, Paris)), Martin Kleinsteuber, Matthias Seibert (TU-Munich),

Theoretical guarantees for 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], [109] [5] on this topic, which concentrated on the noiseless case for non-overcomplete dictionaries, we showed 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 [19]. Moreover, we established sample complexity bounds of dictionary learning and other related matrix factorization schemes (including PCA, NMF, structured sparsity ...) [20].

Learning computationally efficient dictionaries Classical dictionary learning is limited to small-scale problems. Inspired by usual fast transforms, we proposed a general dictionary structure that allows cheaper manipulation, and an algorithm to learn such dictionaries –and their fast implementation. The principle and its application to image denoising appeared at ICASSP 2015 [33] and an application to speedup linear inverse problems was published at EUSIPCO 2015 [32]. A journal paper is currently under revision [51].

We further explored the application of this technique to obtain fast approximations of Graph Fourier Transforms – a conference paper on this latter topic has been accepted for publication in ICASSP 2016 [41]. A C++ software library is in preparation to release the resulting algorithms.

Operator learning for cosparse representations: Besides standard dictionary learning, we also considered learning in the context of the cosparse model. The overall problem is to learn a low-dimensional signal model from a collection of training samples. The mainstream approach is 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.

This year we obtained an upper bound of the sample complexity of the learning process for analysis operators, and designed a stochastic gradient descent (SGD) method to efficiently learn analysis operators with separable structures. Numerical experiments were provided that link the sample complexity to the convergence speed of the SGD algorithm. A journal paper has been published [24].

7.1.3. An alternative framework for sparse representations: analysis sparse models

Participants: Rémi Gribonval, Nancy Bertin, Srdan Kitic, Laurent Albera.

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. Building on our pioneering work on the cosparse model [101], [85], [102] successful applications of this approach to sound source localization, audio declipping and brain imaging have been developed this year.

Versatile co-sparse regularization: Digging the groove of last year results (comparison of the performance of several cosparse recovery algorithms in the context of sound source localization [94], demonstration of its efficiency in situations where usual methods fail ([96], see paragraph 7.5.2), applicability to the hard declipping problem [95], application to EEG brain imaging [60] (see paragraph 7.5.3), a journal paper embedding the latest algorithms and results in sound source localization and brain source localization in a unified fashion was published in IEEE Transactions on Signal Processing [23]. Other communications were made in conferences and workshops [50], [31] and Srdan Kitic defended his PhD thesis [12]. New results include experimental confirmation of robustness and versatility of the proposed scheme, and of its computational merits (convergence speed increasing with the amount of data)

Parametric operator learning for cosparse calibration: In many inverse problems, a key challenge is to cope with unknown physical parameters of the problem such as the speed of sound or the boundary impedance. In the sound source localization problem, we showed that the unknown speed of sound can be learned jointly in the process of cosparse recovery, under mild conditions (work presented last year at iTwist'14 workshop [66]). This year, improved and extended results were obtained: first with a new algorithm for sound source localization with unknown speed of sound [12], then by extending the formulation to the case of unknown boundary impedance, and showing that a similar biconvex formulation and optimization could solve this new problem efficiently (conference paper accepted for publication in ICASSP 2016 [38], see also Section 7.3.2).

7.2. Activities on waveform design for telecommunications

Peak to Average Power Ratio (PAPR), Orthogonal Frequency Division Multiplexing (OFDM), Generalized Waveforms for Multi Carrier (GWMC)

7.2.1. Characterizing multi-carrier waveform systems with optimum PAPR

Participant: Rémi Gribonval.

Main collaboration: Marwa Chafii, Jacques Palicot, Carlos Bader (Equipe SCEE, Supelec, Rennes)

In the context of the TEPN (Towards Energy Proportional Networks) Comin Labs project (see Section 9.1.1.2), in collaboration with the SCEE team at Supelec (thesis of Marwa Chafii co-supervised by R. Gribonval), we investigated a problem related to dictionary design: the characterization of waveforms with low Peak to Average Power Ratio (PAPR) for wireless communications. This is motivated by the importance of a low PAPR for energy-efficient transmission systems. A first stage of the work consisted in characterizing the statistical distribution of the PAPR for a general family of multi-carrier systems, leading to a journal paper [77] and several conference communications [75], [76]. The work this year concentrated on characterizing waveforms with optimum PAPR [30], [48].

7.3. Emerging activities on compressive learning and inverse problems

Compressive sensing, compressive learning, audio inpainting,

7.3.1. Audio inpainting

Participants: Rémi Gribonval, Nancy Bertin, 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.

Building upon our previous pioneering contributions (definition of the audio inpainting problem as a general framework for many audio processing tasks, application to the audio declipping or desaturation problem, formulation as a sparse recovery problem [59]), new results were obtained last year and this year to address the case of audio declipping with the competitive cosparse approach. Last year, its promising results, especially when the clipping level is low, were confirmed experimentally by the formulation and use of a new algorithm named Cosparse Iterative Hard Tresholding [95], which is a counterpart of the sparse Consistent Iterative Hard Thresholding.

This year, we proposed a new algorithmic framework called SPADE, based on non-convex heuristics and which can accomodate both the sparse and cosparse prior. We studied their performance numerically and observed in particular that its cosparse version offers a very appealing trade-off between reconstruction performance and computational time [31], making it suitable for practical applications, even in real-time. We could also confirm our results by subjective listening tests conducted this year [12].

The work on cosparse audio declipping was awarded the Conexant best paper award at the LVA/ICA conference [31] and draw the attention of a world leading company in professional audio signal processing, with which some transfer has been negotiated.

Current and future works deal with developing advanced (co)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, joint source separation and declipping).

7.3.2. Blind Calibration of Impedance and Geometry

Participants: Rémi Gribonval, Nancy Bertin, Srdan Kitic.

Main collaborations: Laurent Daudet, Thibault Nowakowski, Julien de Rosny (Institut Langevin)

This year, we also investigated extended inverse problem scenarios where a "lack of calibration" may occur, i.e., when some physical parameters are needed for reconstruction but apriori unknown: speed of sound, impedance at the boundaries of the domain where the studied phenomenon propagates, or even the shape of these boundaries. In a first approach, based on our physics-driven cosparse regularization of the sound source localization problem [23] (see section 7.1.3), we managed to preserve the sound source localization performance when the speed of sound is unkonwn, or, equally, when the impedance is unknown, provided the shape is and under some smoothness assumptions. Unlike the previous case (gain calibration), the arising problems are not convex but biconvex, and can be solved with proper biconvex formulation of ADMM algorithm. In a second approach based on eigenmode decomposition (limited to a 2D membrane), we showed that impedance learning with known shape, or shape learning with known impedance can be expressed as two facets of the same problem, and solved by the same approach, from a small number of measurements. Two papers presenting these two sets of results were accepted for publication in ICASSP 2016 [38], [35].

7.3.3. Sketching for Large-Scale Mixture Estimation

Participants: Rémi Gribonval, Nicolas Keriven.

Main collaborations: Patrick Perez (Technicolor R&I France) Anthony Bourrier (formerly Technicolor R&I France, now at GIPSA-Lab)

When fitting a probability model to voluminous data, memory and computational time can become prohibitive. In this work, 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 [70], [68], [69]. This year, we consolidated our extensions to non-isotropic Gaussians, with new

algorithms [49] and conducted large-scale experiments demonstrating its potential for speaker verification. A conference paper has been accepted to ICASSP 2016 [40] and a journal version is being finalized.

7.4. Recent results on tensor decompositions

tensor, multiway array, canonical polyadic decomposition, nonnegative tensor factorization

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 dates back to Jordan who was interested in simultaneously diagonalizing two matrices at a time [93]. 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 [112]. 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 [73]. 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 result in a slow convergence (or lack of convergence) of iterative algorithms such as the ALS one [97]. Consequently, a new class of efficient algorithms able to take into account the properties of tensors to be decomposed is needed.

7.4.1. CP decomposition of semi-symmetric three-way arrays subject to arbitrary convex constraints

Participant: Laurent Albera.

Main collaborations : Lu Wang (LTSI, France), Amar Kachenoura (LTSI, France), Lotfi Senhadji (LTSI, France), Jean-Christophe Pesquet (LIGM, France)

We addressed the problem of canonical polyadic decomposition of semi-symmetric 3rd order tensors (i.e. joint diagonalization by congruence) subject to arbitrary convex constraints. Sufficient conditions for the existence of a solution were proved. An efficient algorithm based on the Alternating Direction Method of Multipliers (ADMM) was then designed. ADMM provides an elegant approach for handling the additional constraint terms, while taking advantage of the structure of the objective function. Numerical tests on simulated matrices showed the benefits of the proposed method for low signal to noise ratios. Simulations in the context of nuclear magnetic resonance spectroscopy were also provided. This work was presented at the IEEE CAMSAP'15 conference [29].

7.4.2. Joint eigenvalue decomposition of non-defective matrices for the CP decomposition of tensors

Participant: Laurent Albera.

We proposed a fast and efficient Jacobi-like approach named JET (Joint Eigenvalue decomposition based on Triangular matrices) for the Joint EigenValue Decomposition (JEVD) of a set of real or complex non-defective matrices based on the **LU** factorization of the matrix of eigenvectors [98]. The JEVD can be useful in several contexts such as CP decomposition of tensors [99] and more particularly in Independent Component Analysis (ICA) based on higher order cumulants where it allows us to blindly compute the mixing matrix of sources with kurtosis of different signs. Regarding the proposed JET approach, contrary to classical Jacobi-like JEVD methods, its iterative procedure can be reduced to the search for only one of the two triangular matrices involved in the factorization of the matrix of eigenvectors, hence decreasing the numerical complexity. Two variants of the JET technique, namely JET-U and JET-O, which correspond to the optimization of two different cost functions were described in detail and these were extended to the complex case. Numerical simulations showed that in many practical cases the JET approach provides more accurate estimation of the matrix of eigenvectors than its competitors and that the lowest numerical complexity is consistently achieved by the JET-U algorithm.

7.5. Source separation and localization

Source separation, sparse representations, tensor decompositions, semi-nonnegative independent component analysis, probabilistic model, source localization

Source separation is the task of retrieving the source signals underlying a multichannel mixture signal.

About a decade ago, state-of-the-art approaches consisted of representing the signals in the time-frequency domain and estimating the source coefficients by sparse decomposition in that basis. These approaches rely only on spatial cues, which are often not sufficient to discriminate the sources unambiguously. Over the last years, we proposed a general probabilistic framework for the joint exploitation of spatial and spectral cues [106], which generalizes a number of existing techniques including our former study on spectral GMMs [61]. We showed how it could be used to quickly design new models adapted to the data at hand and estimate its parameters via the EM algorithm, and it became the basis of a large number of works in the field, including our own. In the last years, improvements were obtained through the use of prior knowledge about the source spatial covariance matrices [83], [92], [91], knowledge on the source positions and room characteristics [84], or a better initialization of parameters thanks to specific source localization techniques [67]. This accumulated progress lead to two main achievements last year: a new version of the Flexible Audio Source Separation Toolbox, fully reimplemented, was released [108] and we published an overview paper on recent and going research along the path of *guided* separation, *i.e.*, techniques and models allowing to incorporate knowledge in the process towards efficient and robust solutions to the audio source separation problem, in a special issue of IEEE Signal Processing Magazine devoted to source separation and its applications [113].

7.5.1. Towards real-world separation and remixing applications

Participants: Nancy Bertin, Frédéric Bimbot, Nathan Souviraà-Labastie, Ewen Camberlein, Romain Lebarbenchon.

Main collaboration: Emmanuel Vincent (EPI PAROLE, Inria Nancy)

While some challenges remain, work from previous years and our review paper on guided source separation [113] highlighted that progress has been made and that audio source separation is closer than ever to successful industrial applications, especially when some knowledge can be incorporated. This was exemplified by the contract with MAIA Studio, which reached its end in December 2014 and showed in particular how user input or side information could raise source separation tools to efficient solutions in real-world applications.

In some applicative contexts of source separation, several mixtures are available which contain similar instances of a given source. We have designed a general multi-channel source separation framework where additional audio references are available for one (or more) source(s) of a given mixture. Each audio reference is another mixture which is supposed to contain at least one source similar to one of the target sources. Deformations between the sources of interest and their references are modeled in a linear manner using a generic formulation. This is done by adding transformation matrices to an excitation-filter model, hence affecting different axes, namely frequency, dictionary component or time. A nonnegative matrix co-factorization algorithm

and a generalized expectation-maximization algorithm are used to estimate the parameters of the model. Different model parameterizations and different combinations of algorithms have been tested on music plus voice mixtures guided by music and/or voice references and on professionally-produced music recordings guided by cover references. Our algorithms has provided improvement to the signal-to-distortion ratio (SDR) of the sources with the lowest intensity by 9 to 15 decibels (dB) with respect to the original mixtures [25]. Combining these techniques, with automatic audio motif spotting, we have proposed a new concept called SPORES (for SPOtted Reference based Separation) and applied it to guided separation of audio tracks [13] .

This year saw the beginning of a new industrial collaboration, in the context of the VoiceHome project, aiming at another challenging real-world application: natural language dialog in home applications, such as control of domotic and multimedia devices. As a very noisy and reverberant environment, home is a particularly challenging target for source separation, used here as a pre-processing for speech recognition (and possibly with stronger interactions with voice activity detection or speaker identification tasks as well). In 2015, we participated in a data collection campaign, and in benchmarking and adaptation of existing localization and separation tools to the particular context of this application.

7.5.2. Implicit localization through audio-based control for robotics

Participant: Nancy Bertin.

Main collaborations (audio-based control for robotics): Aly Magassouba and François Chaumette (Inria, EPI LAGADIC, France)

Acoustic source localization is, in general, the problem of determining the spatial coordinates of one or several sound sources based on microphone recordings. This problem arises in many different fields (speech and sound enhancement, speech recognition, acoustic tomography, robotics, aeroacoustics...) and its resolution, beyond an interest in itself, can also be the key preamble to efficient source separation. Common techniques, including beamforming, only provides the *direction of arrival* of the sound, estimated from the *Time Difference of Arrival* (*TDOA*) [67]. This year, we have particularly investigated alternative approaches, either where the explicit localization is not needed (audio-based control of a robot) or, on the contrary, where the exact location of the source is needed and/or TDOA is irrelevant (cosparse modeling of the acoustic field, see Section 7.1.3).

In robotics, the use of aural perception has received recently a growing interest but still remains marginal in comparison to vision. Yet audio sensing is a valid alternative or complement to vision in robotics, for instance in homing tasks. Most existing works are based on the relative localization of a defined system with respect to a sound source, and the control scheme is generally designed separately from the localization system.

In contrast, the approach that we started investigating last year focuses on a sensor-based control approach. We proposed a new line of work, by considering the hearing sense as a direct and real-time input of closed loop control scheme for a robotic task. Thus, and unlike most previous works, this approach does not necessitate any explicit source localization: instead of solving the localization problem, we focus on developing an innovative modeling based on sound features. To address this objective, we placed ourselves in the sensor-based control framework, especially visual servoing (VS) that has been widely studied in the past [78].

From now on, we have established an analytical model linking sound features and control input of the robot, defined and analyzed robotic homing tasks involving multiple sound sources, and validated the proposed approach by simulations and experiments with an actual robot. This work is mainly lead by Aly Magassouba, whose Ph.D. is co-supervised by Nancy Bertin and François Chaumette. A conference paper presenting these first results was published this year [34] and another was submitted to ICRA 2016. Future work will include additional real-world experiments with the robot Romeo from Aldebaran Robotics, investigation of new tasks with active sensing strategies, explicit use of echoes and reverberation to increase robustness, and exploration of dense methods (control from raw acoustic signals rather than from acoustic features).

7.5.3. Brain source localization

Participants: Laurent Albera, Srdan Kitic, Nancy Bertin, Rémi Gribonval.

Main collaborations : Hanna Becker (GIPSA & LTSI, France), Pierre Comon (GIPSA, France), Isabelle Merlet (LTSI, France), Fabrice Wendling (LTSI, France)

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. This work was published in the IEEE Signal Processing Magazine [14].

A sparsity-based approach

Identifying the location and spatial extent of several highly correlated and simultaneously active brain sources from EEG recordings and extracting the corresponding brain signals is a challenging problem. In our comparison of source imaging techniques presented at ICASSP'14 [65], the VB-SCCD algorithm [81], which exploits the sparsity of the variational map of the sources, proved to be a promising approach. We proposed several ways to improve this method. In order to adjust the size of the estimated sources, we added a regularization term that imposes sparsity in the original source domain. Furthermore, we demonstrated the application of ADMM, which permitted to efficiently solve the optimization problem. Finally, we also considered the exploitation of the temporal structure of the data by employing L1,2-norm regularization. The performance of the resulting algorithm, called SISSY, was evaluated based on realistic simulations in comparison to VB-SCCD and several state-of-the-art techniques for extended source localization. This work was partially presented at EUSIPCO'14 [64] and a journal paper is in preparation.

Tensor- and sparsity-based approaches

The separation of EEG sources is a typical application of tensor decompositions in biomedical engineering. The objective of most approaches studied in the literature consists in providing separate spatial maps and time signatures for the identified sources. However, for some applications, a precise localization of each source is required.

To achieve this, a two-step approach was presented at the IEEE EMBC conference [26]. The idea of this approach is to separate the sources using the canonical polyadic decomposition in the first step and to employ the results of the tensor decomposition to estimate distributed sources in the second step, using the SISSY algorithm [64].

Next, we proposed to combine the tensor decomposition and the source localization in a single step [27]. To this end, we directly imposed structural constraints, which are based on a priori information on the possible source locations, on the factor matrix of spatial characteristics. The resulting optimization problem was solved using the alternating direction method of multipliers (ADMM), which was incorporated in the alternating least squares tensor decomposition algorithm. Realistic simulations with epileptic EEG data confirmed that the proposed single-step source localization approach outperformed the previously developed two-step approach.

7.5.4. Independent component analysis

Participant: Laurent Albera.

Main collaboration: Sepideh Hajipour (LTSI & BiSIPL), Isabelle Merlet (LTSI, France), Mohammad Bagher Shamsollahi (BiSIPL, Iran)

Independent Component Analysis (ICA) is a very useful tool to process biomedical signals including EEG data.

We proposed a Jacobi-like Deflationary ICA algorithm, named JDICA. More particularly, while a projection-based deflation scheme inspired by Delfosse and Loubaton's ICA technique (DelL$^{\mathbb{R}}$) [80] was used, a Jacobi-like optimization strategy was proposed in order to maximize a fourth order cumulant-based contrast built from whitened observations. Experimental results obtained from simulated epileptic data mixed with a real muscular activity and from the comparison in terms of performance and numerical complexity with the FastICA [90], RobustICA [114] and DelL$^{\mathbb{R}}$ algorithms, show that the proposed algorithm offers the best trade-off between performance and numerical complexity. This work was published in the IEEE Signal Processing Letters journal [21].

In addition, we illustrated in the ICA context the interest of being able to solve efficiently the (non-orthogonal) JEVD problem. More particularly, we showed that, when the noise covariance matrix is unknown and the source kurtoses have different signs, the joint diagonalization problem involved in the ICAR method [58] becomes a non-orthogonal JEVD problem. Consequently, by using our JET-U algorithm [98], giving birth to the MICAR-U (Modified ICAR based on JET-U) technique, we then provided a more robust ICA method. The identifiability of the MICAR-U technique was studied and proved under some conditions. Computer results given in the context of brain interfaces showed the better ability of the MICAR-U approach to denoise electro-cortical data compared to classical ICA techniques for low signal to noise ratio values. These results were presented in [98].

7.5.5. Semi-nonnegative independent component analysis

Participant: Laurent Albera.

Main collaboration: Lu Wang (LTSI, France), Amar Kachenoura (LTSI, France), Lotfi Senhadji (LTSI, France), Huazhong Shu (LIST, China)

ICA plays also an important role in many other areas including speech and audio [62], [63], [74], [71], radiocommunications [79] and document restoration [111] to cite a few.

For instance in [111], 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 [82] 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 [107], 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 [110]. Moreover, using the JADE algorithm [72] 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 [71] 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 [82], the thickness of each organ, which stands for the mixing coefficient, is real positive. Furthermore, reflectance indices in [111] 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 showed 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 [100]

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 7.4.1 . Hence we use the new method described in section 7.4.1 to perform semi-nonnegative ICA. Performance results in biomedical engineering were given in the paper cited in section 7.4.1 .

7.6. Audio and speech content processing

Audio segmentation, speech recognition, motif discovery, audio mining

7.6.1. Audio motif discovery and spotting

Participants: Frédéric Bimbot, Nathan Souviraà-Labastie.

This work was performed in close collaboration with Emmanuel Vincent from Inria Nancy-Grand Est.

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. When the audio pattern is determined in a user supervised fashion, the task becomes that of motif spotting.

Investigated in the context of SPORES (SPOtted Reference based Separation) [13], audio motif spotting has been illustrated as a useful way to exploit redundancy in audio contents, for guided source separation purposes.

7.6.2. Mobile device for the assistance of users in potentially dangerous situations

Participants: Romain Lebarbenchon, Ewen Camberlein, 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 has been involved in the design of algorithms for the analysis and monitoring of the acoustic scene around the user, yielding audio-based information which can be fused with other sensors (physiological, positional, etc.) in order to trigger an alarm (and subsequent appropriate measures) when needed.

The last phase of the project has been dedicated towards robustness improvement of audio scene analysis, with a particular focus on threat vs non-threat detection, on the basis of adaptive training scenarii. Knowledge and know-how transfer has been achieved for the hardware implementation of the designed methods and the efficient integration into an operational prototype.

7.7. Music Content Processing and Music Information Retrieval

Acoustic modeling, non-negative matrix factorisation, music language modeling, music structure

7.7.1. Music structure modeling by System & Contrast

Participants: Frédéric Bimbot, Corentin Louboutin.

The *System & Contrast* (S&C) model aims at describing the inner organization of structural segments within music pieces in terms of : (i) a carrier system, i.e. a sequence of morphological elements forming a multi-dimensional network of self-deducible syntagmatic relationships and (ii) a contrast, i.e. a substitutive element, usually the last one, which partly departs from the logic implied by the rest of the system [16].

With a primary focus on pop music, the S&C model provides a framework to describe internal implication patterns in musical segments by encoding similarities and relations between its constitutive elements so as to minimize the complexity of the resulting description. It is applicable at several timescales and to a wide variety of musical dimensions in a polymorphous way, therefore offering an attractive meta-description of different types of musical contents.

We have established the filiation of the S&C model as an extension of Narmour's Implication-Realization model [104], [105] and Cognitive Rule-Mapping [103].

We have introduced the Minimum Description Length scheme as a productive paradigm that supports the estimation of S&C descriptions and establishes promising connections between Music Data Processing and Information Retrieval on the one hand, and modern theories in Music Perception and Cognition on the other hand, together with interesting perspectives in other areas in Musicology.

The model is currently being investigated for the multi-scale description of chord sequences.

7.7.2. Tree-based representation of music pieces

Participants: Frédéric Bimbot, Corentin Guichaoua.

Modeling music structure, i.e. the organisation of musical elements and their relationships within a piece of music, is an open problem of primary importance in MIR.

To address this challenge, we approach music structure description as the inference of a low complexity generative grammar able to account for the music piece, itself represented as a sequence of symbols.

Originally introduced for the inference of structure in DNA sequences, Straight-Line Grammars (SLG) form a particular subclass of Context-Free Grammars (CFG) which can be used to model symbolic sequences and to represent them as hierarchical trees. However, SLGs appear to be poorly suited to some particularities of musical patterns, such as segmental regularities, closure substitutions and specific style structures.

We are designing and investigating formal and algorithmic extensions of SLGs as SLEGs (Straight-Line Edition Grammars). Based on a more general minimum description criterion, the SLEG extension allows alterations in the generation step and enables the use of priors in the grammar inference process. Current work includes a diagnostic comparison between the various approaches on the structural segmentation of chord sequences from pop songs.

SIROCCO Project-Team

7. New Results

7.1. 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

7.1.1. Visual attention

Participants: Pierre Buyssens, Olivier Le Meur.

Visual attention is the mechanism allowing to focus our visual processing resources on behaviorally relevant visual information. Two kinds of visual attention exist: one involves eye movements (overt orienting) whereas the other occurs without eye movements (covert orienting). Our research activities deals with the understanding and modeling of overt attention as well as saliency-based image editing. These research activities are described in the following sections.

Saccadic model: Most of the computation models of visual attention output a 2D static saliency map. This single topographic saliency map which encodes the ability of an area to attract our gaze is commonly computed from a set of bottom-up visual features. Although the saliency map representation is a convenient way to indicate where we look within a scene, these models do not completely account for the complexities of our visual system. One obvious limitation concerns the fact that these models do not make any assumption about eye movements and viewing biases. For instance, they implicitly make the hypothesis that eyes are equally likely to move in any direction.

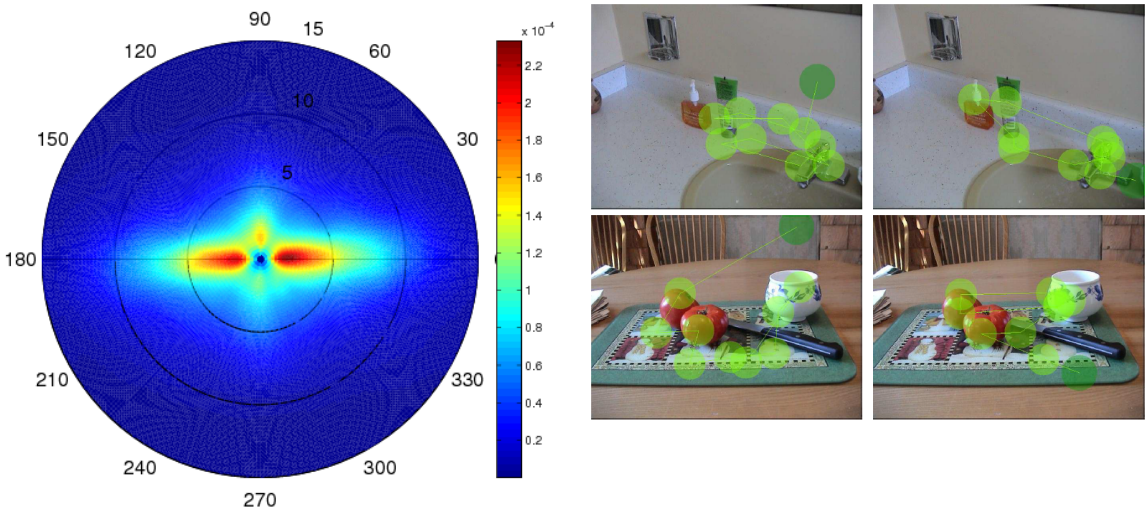
There is evidence for the existence of systematic viewing tendencies. Such biases could be combined with computational models of visual attention in order to better predict where we look. Such a model, predicting the visual scanpath of observer, is termed as saccadic model. We recently propose a saccadic model ([20]) that combines bottom-up saliency maps, viewing tendencies and short-term memory. The viewing tendencies are related to the fact that most saccades are small (less than 3 degrees of visual angle) and oriented in the horizontal direction. Figure 1 (a) illustrates the joint probability distribution of saccade amplitudes and orientations. Examples of predicted scanpaths are shown in Figure 1 (b). We demonstrated that the proposed model outperforms the best state-of-the-art saliency models.

In the future, the goal is to go further by considering that the joint distribution of saccade amplitudes and orientations is spatially variant and depends on the scene category.

Perceptual-based image editing: Since the beginning of October, we have started new studies related to perceptual-based image editing. The goal is to combine the modelling of visual attention with image/video editing methods. More specifically it aims at altering images/video sequences in order to attract viewers attention over specific areas of the visual scene. We intend to design new computational editing methods for emphasizing and optimizing the importance of pre-defined areas of the input image/video sequence. There exist very few studies in the literature dealing with this problem. Current methods simply alter the content by using blurring operation or by recoloring the image locally so that the focus of attention falls within the pre-defined areas of interest. One avenue for improving current methods is to minimize a distance computed between a user's defined visual scanpath and predicted visual scanpath. The content would be edited (i.e. recoloring, region rescaling, local contrast/resolution adjustment, removing disturbing object, etc) in an iterative manner in order to move the focus of attention towards the regions selected by the user.

7.1.2. Epitome-based video representation

Participants: Martin Alain, Christine Guillemot.



(a) (b)

Figure 1. (a) Joint probability distribution of saccade amplitudes and orientations shown on a polar plot. Radial position indicates saccadic amplitudes expressed in degree of visual angle. (b) Predicted scanpaths composed of ten fixations represented by green circles. The dark green circle corresponds to the first fixation which is randomly chosen.

In 2014, we have developed fast methods for constructing epitomes from images. An epitome is a factorized texture representation of the input image, and its construction exploits self-similarities within the image. Known construction methods are memory and time consuming. The proposed methods, using dedicated list construction on one hand and clustering techniques on the other hand, aim at reducing the complexity of the search for self-similarities.

In 2015, we have developed methods for quantization noise removal (after decoding) exploiting the epitome representations together with local learning of either LLE (locally linear embedding) weights, which has proved to be a powerful tool for prediction [14], or using linear mapping functions between original and noisy patches. Compared to classical denoising methods which, most of the time, assume additive white Gaussian noise, the quantization turns out to be correlated to the signal which makes the problem more difficult. The methods have been experimented both in the contexts of single layer encoding and scalable encoding. The same methodology has been applied to super-resolution learning this time mapping functions between the low resolution and high resolution spaces in which lie the patches of the epitome [32].

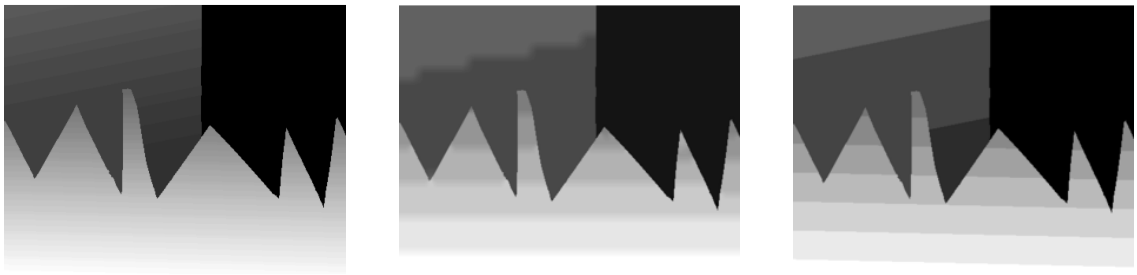
7.1.3. Graph-based multi-view video representation

Participants: Christine Guillemot, Thomas Maugey, Mira Rizkallah, Xin Su.

One of the main open questions in multiview data processing is the design of representation methods for multiview data, where the challenge is to describe the scene content in a compact form that is robust to lossy data compression. Many approaches have been studied in the literature, such as the multiview and multiview plus depth formats, point clouds or mesh-based techniques. All these representations contain two types of data: i) the color or luminance information, which is classically described by 2D images; ii) the geometry information that describes the scene 3D characteristics, represented by 3D coordinates, depth maps or disparity vectors. Effective representation, coding and processing of multiview data partly rely on a proper representation of the geometry information. The multiview plus depth (MVD) format has become very popular in recent years for 3D data representation. However, this format induces very large volumes of data, hence the

need for efficient compression schemes. On the other hand, lossy compression of depth information in general leads to annoying rendering artefacts especially along the contours of objects in the scene.

Instead of lossy compression of depth maps, we consider the lossless transmission of a geometry representation that captures only the information needed for the required view reconstructions. Our goal is to transmit "just enough" geometry information for accurate representation of a given set of views, and hence better control the effect of geometry lossy compression.



(a) (b) (c)

Figure 2. (a) original depth map, (b) depth map compressed with edge-adaptive method at 10kb with compression artifacts (c) depth image retrieved from the graph of our proposed GBR transmitted at 10kb keeping the original scene structure.

More particularly, in [23], we proposed a new Graph-Based Representation (GBR) for geometry information, where the geometry of the scene is represented as connections between corresponding pixels in different views. In this representation, two connected pixels are neighboring points in the 3D scene. The graph connections are derived from dense disparity maps and provide just enough geometry information to predict pixels in all the views that have to be synthesized.

GBR drastically simplifies the geometry information to the bare minimum required for view prediction. This "task-aware" geometry simplification allows us to control the view prediction accuracy before coding compared to baseline depth compression methods (Fig. 2). This work has first been carried out for multi-view configurations, in which cameras are parallel. We are currently investigating the extension of this promising GBR to complex camera transitions. An algorithm has already been implemented for two views and is being extended for multiple views. The next steps will be to develop color coding tools adapted to these graph structures.

7.2. Rendering, inpainting and super-resolution

image-based rendering, inpainting, view synthesis, super-resolution

7.2.1. Color and light transfer

Participants: Hristina Hristova, Olivier Le Meur.

Color transfer aims at modifying the look of an original image considering the illumination and the color palette of a reference image. It can be employed for image and video enhancement by simulating the appearance of a given image or a video sequence. It can also be applied to hallucinations of particular parts of the day. Current state-of-the-art methods focus mainly on the global transfer of the light and color distributions. Unfortunately, the use of a global distribution is questionable since the light and color of image can significantly vary within the same scene. In [27], we proposed a new method to deal with the limitations of existing methods. The proposed method aims at performing the partitions of the input and reference images

into Gaussian distributed clusters by considering the main style of input and reference images. From this clustering, several novel policies are defined for mapping the clusters of the input and reference images. To complete the style transfer, for each pair of corresponding clusters, we apply a parametric color transfer method (i.e. Monge-Kantorovitch transformation) and a local chromatic adaptation transform. Results, subjective user evaluation as well as objective evaluation show that the proposed method obtains visually pleasing and artifact-free images, respecting the reference style. Some results are illustrated in Figure 3 .

Figure 3. From left to right: input image, reference image and the result of the proposed method.

In [34], we extended the method presented in [27] to deal with a color transfer between two HDR images. One limitation of the two proposed methods is that we are still considering that the distributions of color and light follow a Gaussian law. We are currently investigating a more general approach by considering multivariate generalized Gaussian distribution.

7.2.2. Image guided inpainting

Participants: Christine Guillemot, Thomas Maugey.

Inpainting of images has been intensively studied in the past few years, especially for applications such as image restoration and editing [16]. Another application where inpainting techniques are useful is view synthesis, where holes are to be filled corresponding to areas that are no longer occluded. In the particular cases where one has access to ground truth images (like for example in multiview video coding where view synthesis is used for predicting the captured views from a reference one), auxiliary information can be generated to help inpainting, which leads to the concept of *guided inpainting*.

In [29], we have proposed a new auxiliary information that is used to refine the set of candidate patches for the hole filling step of the inpainting. Assuming that the patches of an image lie in a union of subspaces, *i.e.*, the images have different regions with different color textures, these patches are first clustered using a new recursive spectral clustering algorithm that extends existing sparse subspace clustering and replaces the sparse approximation by locally linear embedding, better suited for the inpainting context. Dictionaries are then built from these clusters and used for the hole filling process. However, the inpainting is not always able to "guess" in which cluster the patches of the hole belong to (especially around discontinuities). The auxiliary information that is built from the ground truth image may help to find the right cluster. We thus propose a new guided inpainting algorithm that forces the patch reconstruction to be done in one cluster only, if no auxiliary information is available, or in the cluster pointed by the auxiliary information, if it is available. Experiments (Fig. 4) show that auxiliary information helps to significantly improve the inpainting quality for a reasonable coding cost.

We are currently working on the extension of this technique in order to place the guided inpainting problem in an information theoretic framework, and better answer the following questions: when additional information is actually needed? What type of auxiliary information is needed? how to optimize in a rate-distortion sense the guided inpainting problem?.

(a) (b) (c)

Figure 4. (a) input image to inpaint, (b) filled image using baseline not guided inpainting (c) filled image using proposed guided inpainting with an auxiliary information cost of 0.018 *bpp bitrate.*

7.2.3. Clustering on manifolds for image restoration

Participants: Julio Cesar Ferreira, Christine Guillemot, Elif Vural.

Local learning of sparse image models has proven to be very effective to solve a variety of inverse problems in many computer vision applications. To learn such models, the data samples are often clustered using the K-means algorithm with the Euclidean distance as a dissimilarity metric. However, the Euclidean distance may not always be a good dissimilarity measure for comparing data samples lying on a manifold. We have developed two algorithms for determining a local subset of training samples from which a good local model can be computed for reconstructing a given input test sample, where we take into account the underlying geometry of the data. The first algorithm, called Adaptive Geometry-driven Nearest Neighbor search (AGNN), is an adaptive scheme which can be seen as an out-of-sample extension of the replicator graph clustering method for local model learning. The second method, called Geometry-driven Overlapping Clusters (GOC), is a less complex nonadaptive alternative for training subset selection. The AGNN and GOC methods have been evaluated in image super-resolution, deblurring and denoising applications and shown to outperform spectral clustering, soft clustering, and geodesic distance based subset selection in most settings.

7.3. Representation and compression of large volumes of visual data

Sparse representations, data dimensionality reduction, compression, scalability, perceptual coding, rate-distortion theory

7.3.1. Manifold learning and low dimensional embedding for classification

Participants: Christine Guillemot, Elif Vural.

Typical supervised classifiers such as SVM are designed for generic data types and do not make any particular assumption about the geometric structure of data, while data samples have an intrinsically low-dimensional structure in many data analysis applications. Recently, many supervised manifold learning methods have been proposed in order to take the low-dimensional structure of data into account when learning a classifier. Unlike unsupervised manifold learning methods which only take the geometric structure of data samples into account when learning a low-dimensional representation, supervised manifold learning methods learn an embedding that not only preserves the manifold structure in each class, but also enhances the separation between different classes.

An important factor that influences the performance of classification is the separability of different classes in the computed embedding. We have done a theoretical analysis of separability of data representations given by supervised manifold learning. In particular, we have focused on the nonlinear supervised extensions of the Laplacian eigenmaps algorithm and have examined the linear separation between different classes in the learned embedding. We have shown that, if the graph is such that the inter-group graph weights are sufficiently small, the learned embedding becomes linearly separable at a dimension that is proportional to the number of groups. These theoretical findings have been confirmed by experimentation on synthetic data sets and image data.

We have then considered the problem of out-of-sample generalizations for manifold learning. Most manifold learning methods compute an embedding in a pointwise manner, i.e., data coordinates in the learned domain are computed only for the initially available training data. The generalization of the embedding to novel data samples is an important problem, especially in classification problems. Previous works for out-of-sample generalizations have been designed for unsupervised methods. We have studied this problem for the particular application of data classification and proposed an algorithm to compute a continuous function from the original data space to the low-dimensional space of embedding. In particular, we have constructed an interpolation function in the form of a radial basis function that maps input points as close as possible to their projections onto the manifolds of their own class. Experimental results have shown that the proposed method gives promising results in the classification of low-dimensional image data such as face images.

7.3.2. Adaptive clustering with Kohonen self-organizing maps for second-order prediction

Participants: Christine Guillemot, Bihong Huang.

The High Efficiency Video Coding standard (HEVC) supports a total of 35 intra prediction modes which aim at reducing spatial redundancy by exploiting pixel correlation within a local neighborhood. However the correlation remains in the residual signals of intra prediction, leading to some high energy prediction residuals. In 2014, we have studied several methods to exploit remaining correlation in residual domain after intra prediction. These methods are based on vector quantization with codebooks learned and dedicated to the different prediction modes in order to model the directional characteristics of the residual signals. The best matching code vector is found in a rate-distortion optimization sense. Finally, the index of the best matching code vector is sent to the decoder and the vector quantization error, the difference between the intra residual vector and the best matching code vector, is processed by the conventional operations of transform, scalar quantization and entropy coding.

In a first approach called MDVQ (Mode Dependent Vector Quantization), the codebooks were learned using the k-means algorithm [26]. More recently, we have developed a variant of the approach, called AMDVQ (Adaptive MDVQ) by adding a codebook update step based on Kohonen Self-Organized Maps which aims at capturing the variations of the residual signal statistical charateristics. The Kohonen algorithm uses previously reconstructed residual vectors to continuously update the code vectors during the encoding and decoding of the video sequence [12].

7.3.3. Rate-distortion optimized tone curves for HDR video compression

Participants: David Gommelet, Christine Guillemot, Aline Roumy.

High Dynamic Range (HDR) images contain more intensity levels than traditional image formats. Instead of 8 or 10 bit integers, floating point values requiring much higher precision are used to represent the pixel data. These data thus need specific compression algorithms. In collaboration with Envivio, we have developed a novel compression algorithm that allows compatibility with the existing Low Dynamic Range (LDR) broadcast architecture in terms of display, compression algorithm and datarate, while delivering full HDR data to the users equipped with HDR display. The developed algorithm is thus a scalable video compression offering a base layer that corresponds to the LDR data and an enhancement layer, which together with the base layer corresponds to the HDR data. The novelty of the approach relies on the optimization of a mapping called Tone Mapping Operator (TMO) that maps efficiently the HDR data to the LDR data. The optimization has been carried out in a rate-distortion sense: the distortion of the HDR data is minimized under the constraint of

minimum sum datarate (for the base and enhancement layer), while offering LDR data that are close to some "aesthetic" a priori. Taking into account the aesthetic of the scene in video compression is novel, since video compression is traditionally optimized to deliver the smallest distortion with the input data at the minimum datarate.

7.3.4. Local Inverse Tone Curve Learning for HDR Image Scalable Compression

Participants: Christine Guillemot, Mikael Le Pendu.

In collaboration with Technicolor, we have developed local inverse tone mapping operators for scalable high dynamic range (HDR) image coding. The base layer is a low dynamic range (LDR) version of the image that may have been generated by an arbitrary Tone Mapping Operator (TMO). No restriction is imposed on the TMO, which can be either global or local, so as to fully respect the artistic intent of the producer. The method which has been developed successfully handles the case of complex local TMOs thanks to a block-wise and non-linear approach [28]. A novel template based Inter Layer Prediction (ILP) is designed in order to perform the inverse tone mapping of a block without the need to transmit any additional parameter to the decoder. This method enables the use of a more accurate inverse tone mapping model than the simple linear regression commonly used for blockwise ILP [21]. In addition, this paper shows that a linear adjustment of the initially predicted block can further improve the overall coding performance by using an efficient encoding scheme of the scaling parameters. Our experiments have shown an average bitrate saving of 47% on the HDR enhancement layer, compared to previous local ILP methods.

7.3.5. HEVC-based UHD video coding optimization

Participants: Nicolas Dhollande, Christine Guillemot, Olivier Le Meur.

The HEVC (High Efficiency Video Coding) standard brings the necessary quality versus rate performance for efficient transmission of Ultra High Definition formats (UHD). However, one of the remaining barriers to its adoption for UHD content is the high encoding complexity. We address the reduction of HEVC encoding complexity by investigating different strategies: First we have proposed to infer UHD coding modes and quad-tree from a first encoding pass which consists in encoding a lower resolution version of the input video. In the context of our study, the first encoding pass encodes a HD video sequence. A speed-up by a factor of 3 is achieved compared to directly encoding the UHD format without compromising the final video quality. The second strategy focuses on the block partitioning of intra frame coding. The Coding Tree Unit (CTU) is the root of the coding tree and can be recursively split into four square Coding Unit (CU), given that the smallest block size is 8×8. Once the partitioning procedure is fully completed, the final quad-tree can be obtained by choosing the configuration leading to the best rate-distortion trade-off. Rather than performing an exhaustive partitioning, we aim to predict the quad-tree partition into coding units (CU). This prediction is based on low-level visual features extracted from the video sequences. The low-level features are related to gradient-based statistics, structure tensors statistics or entropy etc. From these features, we trained a probabilistic model on a set of UHD training sequences in order to determine whether the coding unit should be further split or not. The proposed methods yield a significant encoder speed-up ratio (up to 5.3 times faster) with a moderate loss in terms of compression efficiency [33].

7.4. 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

7.4.1. Information theoretical bounds of Free-viewpoint TV

Participants: Thomas Maugey, Aline Roumy.

Free-viewpoint television FTV is a new system for viewing video where the user can choose its viewpoint freely and change it at anytime. The goal is to propose an immersive sensation without the disadvantage of Three-dimensional (3D) television (TV). Indeed, the conventional 3D displays (with or without glasses) occur, by construction, an accommodation-vergence conflict: since the eye tend to focus on the display screen (accommodation), whereas the brain perceives the depth of 3D images due to the different views seen by each eye (vergence). Instead, with FTV, a look-around effect is produced without any visual fatigue since the displayed images remain 2D. Therefore, FTV presents nice properties that makes it a serious competitor for 3DTV. Existing compression algorithms for FTV consider to send all the views, which would require about 100 Mbits/s (for 100 views, as needed to propose a true navigation within the scene). Since this amount does not fit the current datarate for transmission in a streaming scenario, we investigate a solution where the server only send the request. In [31], [30], we have shown a very surprising and positive result: if all the views are compressed once and if the server extracts from the compressed bitstream the request (i.e. one view at a time), the datarate is exactly the same as if the whole database was entirely decoded, and the requested views reencoded. This very positive result shows that it is possible to send FTV with the same datarate as single view television with very limited computational cost at the server (only extraction from the bistream). This result is an information theoretical result and the goal is now to build a practical system that can achieve this performance.

7.4.2. Compressed Sensing : a probabilistic analysis of the orthogonal matching pursuit algorithm

Participant: Aline Roumy.

Compressed sensing (CS) is an efficient acquisition scheme, where the data are projected onto a randomly chosen subspace to achieve data dimensionality reduction. The projected data are called measurements. The reconstruction is performed from these measurements, by solving underdetermined linear systems under a sparsity a priori constraint. It is generally believed that the greedy algorithm Orthogonal Matching pursuit performs well and can determine which variables are active (i.e. non zero). In contrast, we showed that this is not the case even in the noiseless context. We derived an exact probabilistic analysis of the iterative algorithm in the large system regime, when all dimensions tend to infinity. We showed that as the number of iterations grows, the algorithm will make errors with probability one.