RESEARCH CENTER
Rennes - Bretagne-Atlantique

FIELD

Activity Report 2017

Section New Results

Edition: 2018-02-19

CAIRN Project-Team

7. New Results

7.1. Reconfigurable Architecture Design

7.1.1. Voltage Over-Scaling for Error-Resilient Applications

Participants: Rengarajan Ragavan, Benjamin Barrois, Cédric Killian, Olivier Sentieys.

Voltage scaling has been used as a prominent technique to improve energy efficiency in digital systems, scaling down supply voltage effects in quadratic reduction in energy consumption of the system. Reducing supply voltage induces timing errors in the system that are corrected through additional error detection and correction circuits. In [43], we proposed voltage over-scaling based approximate operators for applications that can tolerate errors. We characterized the basic arithmetic operators using different operating triads (combination of supply voltage, body-biasing scheme and clock frequency) to generate models for approximate operators. Error-resilient applications can be mapped with the generated approximate operator models to achieve optimum trade-off between energy efficiency and error margin. Based on the dynamic speculation technique, best possible operating triad is chosen at runtime based on the user definable error tolerance margin of the application. In our experiments in 28nm FDSOI, we achieved maximum energy efficiency of 89% for basic operators like 8-bit and 16-bit adders at the cost of 20% Bit Error Rate (ratio of faulty bits over total bits) by operating them in near-threshold regime.

7.1.2. Stochastic Computation Elements with Correlated Input Streams

Participants: Rengarajan Ragavan, Rahul Kumar Budhwani, Olivier Sentieys.

In recent years, shrinking size in integrated circuits has imposed a big challenge in maintaining the reliability in conventional computing. Stochastic Computing (SC) has been seen as a reliable, low-cost, and low-power alternative to overcome such issues. SC computes data in the form of bit streams of 1s and 0s. Therefore, SC outperforms conventional computing in terms of tolerance to soft error and uncertainty at the cost of increased computational time. Stochastic Computing with uncorrelated input streams requires streams to be highly independent for better accuracy. This results in more hardware consumption for conversion of binary numbers to stochastic streams. Correlation can be used to design Stochastic Computation Elements (SCE) with correlated input streams. These designs have higher accuracy and less hardware consumption. In [38], we proposed new SC designs to implement image processing algorithms with correlated input streams. Experimental results of proposed SC with correlated input streams show on average 37% improvement in accuracy with reduction of 50-90% in area and 20-85% in delay over existing stochastic designs.

7.1.3. Fault Tolerant Architectures

Participants: Olivier Sentieys, Angeliki Kritikakou, Rafail Psiakis.

Error occurrence in embedded systems has significantly increased, whereas critical applications require reliable processors that combine performance with low cost and energy consumption. Very Long Instruction Word (VLIW) processors have inherent resource redundancy which is not constantly used due to application's fluctuating Instruction Level Parallelism (ILP). Approaches can benefit these additional resources to provide fault tolerance.

The reliability through idle slots utilization can be explored either at compile-time, increasing code size and storage requirements, or at run-time only inside the current instruction bundle, adding unnecessary time slots and degrading performance. To address this issue, we proposed a technique in [41] to explore the idle slots inside and across original and replicated instruction bundles reclaiming more efficiently the idle slots and creating a compact schedule. To achieve this, a dependency analysis is applied at run-time. The execution of both original and replicated instructions is allowed at any adequate function unit, providing higher flexibility on instruction scheduling. The proposed technique achieves up to 26% reduction in performance degradation over existing approaches.

When permanent and soft errors coexist, spare units have to be used or the executed program has to be modified through self-repair or by using several stored versions. However, these solutions introduce high area overhead for the additional resources, time overhead for the execution of the repair algorithm and storage overhead of the multi-versioning. To address these limitations, a hardware mechanism is proposed in [42] which at run-time replicates the instructions and schedules them at the idle slots considering the resource constraints. If a resource becomes faulty, the proposed approach efficiently rebinds both the original and replicated instructions during execution. In this way, the area overhead is reduced, as no spare resources are used, whereas time and storage overhead are not required. Results show up to 49% performance gain over existing techniques.

7.1.4. Hardware Accelerated Simulation of Heterogeneous Platforms

Participants: Minh Thanh Cong, François Charot, Steven Derrien.

When considering designing heterogeneous multi-core platforms, the number of possible design combinations leads to a huge design space, with subtle trade-offs and design interactions. To reason about what design is best for a given target application requires detailed simulation of many different possible solutions. Simulation frameworks exist (such as gem5) and are commonly used to carry out these simulations. Unfortunately, these are purely software-based approaches and they do not allow a real exploration of the design space. Moreover, they do not really support highly heterogeneous multi-core architectures. These limitations motivate the study of the use of hardware to accelerate the simulation, and in particular of FPGA components. In this context, we are currently investigating the possibility of building hardware accelerated simulators using the HAsim simulation infrastructure, jointly developed by MIT and Intel. HAsim is an FPGA-accelerated simulator that is able to simulate a multicore with a high-detailed pipeline, cache hierarchy and detailed on-chip network on a single FPGA. A model of the RISC-V instruction set architecture suited to the HAsim infrastructure has been developed, its deployment on the Xeon+FPGA Intel platform is in progress. This work is done with the perspective of studying hardware accelerated simulation of heterogeneous multicore architectures mixing RISC-V cores and hardware accelerators.

7.1.5. Optical Interconnections for 3D Multiprocessor Architectures

Participants: Jiating Luo, Ashraf El-Antably, Van Dung Pham, 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 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 [37] we designed an Optical-Network-Interface (ONI) to connect a cluster of several processors to the optical communication medium. This interface, constrained by the 10 Gb/s data-rate of the Lasers, integrates Error Correcting Codes (ECC) and a communication manager. This manager can select, at run-time, the communication mode to use depending on timing or power constraints. Indeed, as the use of ECC is based on redundant bits, it increases the transmission time, but saves power for a given Bit Error Rate (BER). Moreover, our ONI allows for data to be sent using several wavelengths in parallel, hence increasing transmission bandwidth. From the design of this interface, estimation in terms of power consumption and execution time have been obtained, as well as the energy per bit of each communication.

The optical medium can support multiple transactions at the same time on different wavelengths by using Wavelength Division Multiplexing (WDM). Moreover, multiple wavelengths can be gathered as high-bandwidth channel to reduce transmission time. However, multiple signals sharing simultaneously a waveguide lead to inter-channel crosstalk noise. This problem impacts the Signal to Noise Ratio (SNR) of the optical signal, which increases the Bit Error Rate (BER) at the receiver side. In [39], we formulated the crosstalk noise and execution time models and then proposed a Wavelength Allocation (WA) method in a ring-based WDM ONoC to reach performance and energy trade-offs based on the application constraints. We showed that for a 16-core ONoC architecture using 12 wavelengths, more than 10^5 allocation solutions exist and only 51 are on a Pareto front giving a tradeoff between execution time and energy per bit (derived from the BER). These optimized solutions reduce the execution time by 37% or the energy from 7.6fJ/bit to 4.4fJ/bit.

We also proposed to explore the selection of laser power for each communication. This approach reduces the global power consumption by ensuring the targeted Bit Error Rate for each communication. To support laser power selection, we have also studied, designed and evaluated at transistor level different configurable laser drivers using a 28NM FDSOI technology.

7.1.6. Adaptive Dynamic Compilation for Low-Power Embedded Systems

Participants: Steven Derrien, Simon Rokicki.

Single ISA-Heterogeneous multi-cores such as the ARM big.LITTLE have proven to be an attractive solution to explore different energy/performance trade-offs. Such architectures combine Out of Order cores with smaller in-order ones to offer different power/energy profiles. They however do not really exploit the characteristics of workloads (compute-intensive vs. control dominated).

In this work, we propose to enrich these architectures VLIW cores, which are very efficient at compute-intensive kernels. To preserve the single ISA programming model, we resort to Dynamic Binary Translation as used in Transmeta Crusoe and NVidia Denver processors. Our proposed DBT framework targets the RISC-V ISA, for which both OoO and in-order implementations exist.

Since DBT operates at runtime, its execution time is directly perceptible by the user, hence severely constrained. As a matter of fact, this overhead has often been reported to have a huge impact on actual performance, and is considered as being the main weakness of DBT based solutions. This is particularly true when targeting a VLIW processor: the quality of the generated code depends on efficient scheduling; unfortunately scheduling is known to be the most time-consuming component of a JIT compiler or DBT. Improving the responsiveness of such DBT systems is therefore a key research challenge. This is however made very difficult by the lack of open research tools or platform to experiment with such platforms.

To address these issues, we have developed an open hardware/software platform supporting DBT. The platform was designed using HLS tools and validated on a FPGA board. The DBT uses RISC-V as host ISA, and can be retargeted to different VLIW configurations. Our platform uses custom hardware accelerators to improve the reactivity of our optimizing DBT flow. Our results [44] show that, compared to a software implementation, our approach offers speed-up by 8× while consuming 18× less energy.

Our current research work investigates how DBT techniques can be used to support runtime configurable VLIW cores. Such cores enable fine grain exploration of energy/performance trade-off by dynamically adjusting their number of execution slots, their register file size, etc.). More precisely, we build on our DBT framework to enable dynamic code specialization. Our first experimental results suggest that this approach leads to best-case performance and energy efficiency when compared against static VLIW configurations [54].

7.1.7. Design Space Exploration for Iterative Stencil computations on FPGA accelerators

Participants: Steven Derrien, Gaël Deest, Tomofumi Yuki.

Iterative stencil computations arise in many application domains, ranging from medical imaging to numerical simulation. Since they are computationally demanding, a large body of work addressed the problem of parallelizing and optimizing stencils for multi-cores, GPUs, and FPGAs. Earlier attempts targeting FPGAs showed that the performance of such accelerators is the result of a complex interplay between the FPGA's raw computing power, the amount of on-chip memory it has, and the performance of the external memory system. They also illustrate how each application may have different requirements. For example, in the context of embedded vision, the designer's goal is often to find the design with minimum cost that matches real-time performance constraints (e.g., 4K@60fps). In an exascale context, the designer's goal is to maximize performance (measured in ops-per-second) for a given FPGA board, while maintaining power dissipation to a minimum. Based on these observations, we explore a family of design options that can accommodate a large set of requirements and constraints, by exposing trade-offs between computing power, bandwidth requirements, and FPGA resource usage. We have developed a code generator that produces HLS-optimized C/C++ descriptions of accelerator instances targeting emerging System on Chip platforms, (e.g., Xilinx Zynq or Intel SoC). Our family of designs builds upon the well-known tiling transformation, which we use to balance on-chip memory cost and off-chip bandwidth. To ease the exploration of this design space, we propose performance models to hone in on the most interesting design points, and show how they accurately lead to optimal designs. Our results demonstrate that the optimal choice depends on problem sizes and performance goals [30].

7.1.8. Energy-driven Accelerator Exploration for Heterogeneous Multiprocessor Architectures

Participants: Baptiste Roux, Olivier Sentieys.

Programming heterogeneous multiprocessor architectures combining multiple processor cores and hardware accelerators is a real challenge. Computer-aided design and development tools try to reduce the large design space by simplifying hardware software mapping mechanisms. However, energy consumption is not well supported in most of design space exploration methodologies due to the difficulty to fast and accurately estimate energy consumption. To this aim,we proposed and validated an exploration method for partitioning applications on software cores and hardware accelerators under energy-efficiency constraints. The methodology is based on energy and performance measurement of a tiny subset of the design space and an analytical formulation of the performance and energy of an application kernel mapped on a heterogeneous architecture. This closed-form expression is captured and solved using Mixed Integer Linear Programming, which allows for very fast exploration resulting in the optimal solution. The approach is validated on two applications kernels using Zynq-based architecture showing more than 12% acceleration speed-up and energy saving compared to standard approaches. Results also show that the most energy-efficient solution is application- and platform-dependent and moreover hardly predictable, which highlights the need for fast exploration.

7.2. Compilation and Synthesis for Reconfigurable Platform

7.2.1. Superword-Level Parallelism-Aware Word Length Optimization

Participants: Steven Derrien, Ali Hassan El Moussawi.

Many embedded processors do not support floating-point arithmetic in order to comply with strict cost and power consumption constraints. But, they generally provide support for SIMD as a mean to improve performance for little cost overhead. Achieving good performance when targeting such processors requires the use of fixed-point arithmetic and efficient exploitation of SIMD data-path. To reduce time-to-market, automatic SIMDization – such as superword level parallelism (SLP) extraction – and floating-point to fixed-point conversion methodologies have been proposed. In [33], we showed that applying these transformations independently is not efficient. We proposed an SLP-aware word length optimization algorithm to jointly perform floating-point to fixed-point conversion and SLP extraction. We implemented the proposed approach

in a source-to-source compiler framework and evaluated it on several embedded processors. Experimental results illustrated the validity of our approach with performance improvement by up to 40% for a limited loss in accuracy.

7.2.2. Automatic Parallelization Techniques for Time-Critical Systems

Participants: Steven Derrien, Imen Fassi, Thomas Lefeuvre.

Real-time systems are ubiquitous, and many of them play an important role in our daily life. In hard real-time systems, computing the correct results is not the only requirement. In addition, the results must be produced within pre-determined timing constraints, typically deadlines. To obtain strong guarantees on the system temporal behavior, designers must compute upper bounds of the Worst-Case Execution Times (WCET) of the tasks composing the system. WCET analysis is confronted with two challenges: (i) extracting knowledge of the execution flow of an application from its machine code, and (ii) modeling the temporal behavior of the target platform. Multi-core platforms make the latter issue even more challenging, as interference caused by concurrent accesses to shared resources have also to be modeled. Accurate WCET analysis is facilitated by *predictable* hardware architectures. For example, platforms using ScratchPad Memories (SPMs) instead of caches are considered as more predictable. However SPM management is left to the programmer-managed, making them very difficult to use, especially when combined with complex loop transformations needed to enable task level parallelization. Many researches have studied how to combine automatic SPM management with loop parallelization at the compiler level.It has been shown that impressive average-case performance improvements could be obtained on compute intensive kernels, but their ability to reduce WCET estimates remains to be demonstrated, as the transformed code does not lends itself well to WCET analysis.

In the context of the ARGO project, and in collaboration with members of the PACAP team, we have studied how parallelizing compilers techniques should be revisited in order to help WCET analysis tools. More precisely, we have demonstrated the ability of polyhedral optimization techniques to reduce WCET estimates in the case of sequential codes, with a focus on locality improvement and array contraction. We have shown on representative real-time image processing use cases that they could bring significant improvements of WCET estimates (up to 40%) provided that the WCET analysis process is guided with automatically generated flow annotations [31].

7.2.3. Operator-Level Approximate Computing

Participants: Benjamin Barrois, Olivier Sentieys.

Many applications are error-resilient, allowing for the introduction of approximations in the calculations, as long as a certain accuracy target is met. Traditionally, fixed-point arithmetic is used to relax accuracy, by optimizing the bit-width. This arithmetic leads to important benefits in terms of delay, power and area. Lately, several hardware approximate operators were invented, seeking the same performance benefits. However, a fair comparison between the usage of this new class of operators and classical fixed-point arithmetic with careful truncation or rounding, has never been performed. In [27], we first compare approximate and fixed-point arithmetic operators in terms of power, area and delay, as well as in terms of induced error, using many state-of-the-art metrics and by emphasizing the issue of data sizing. To perform this analysis, we developed a design exploration framework, *ApxPerf*, which guarantees that all operators are compared using the same operating conditions. Moreover, operators are compared in several classical real-life applications leveraging relevant metrics. In [27], we show that considering a large set of parameters, existing approximate adders and multipliers tend to be dominated by truncated or rounded fixed-point ones. For a given accuracy level and when considering the whole computation data-path, fixed-point operators are several orders of magnitude more accurate while spending less energy to execute the application. A conclusion of this study is that the entropy of careful sizing is always lower than approximate operators, since it require significantly less bits to be processed in the data-path and stored. Approximated data therefore always contain on average a greater amount of costly erroneous, useless information.

In [26] we performed a comparison between custom fixed-point (FxP) and floating-point (FlP) arithmetic, applied to bidimensional K-means clustering algorithm. First, FxP and FlP arithmetic operators are compared in terms of area, delay and energy, for different bitwidth, using the *ApxPerf2.0* framework. Finally, both are compared in the context of K-means clustering. The direct comparison shows the large difference between 8-to-16-bit FxP and FlP operators, FlP adders consuming 5-12$\times$ more energy than FxP adders, and multipliers 2-10$\times$ more. However, when applied to K-means clustering algorithm, the gap between FxP and FlP tightens. Indeed, the accuracy improvements brought by FlP make the computation more accurate and lead to an accuracy equivalent to FxP with less iterations of the algorithm, proportionally reducing the global energy spent. The 8-bit version of the algorithm becomes more profitable using FlP, which is 80% more accurate with only $1.6\times$ more energy.

7.2.4. Dynamic Fault-Tolerant Mapping and Scheduling on Multi-core systems

Participants: Emmanuel Casseau, Petr Dobias.

Demand on multi-processor systems for high performance and low energy consumption still increases in order to satisfy our requirements to perform more and more complex computations. Moreover, the transistor size gets smaller and their operating voltage is lower, which goes hand in glove with higher susceptibility to system failure. In order to ensure system functionality, it is necessary to conceive fault-tolerant systems. One way to tackle this issue is to makes use of both the redundancy and reconfigurable computing, especially when multi-processor platforms are targeted. Actually, multi-processor platforms can be less vulnerable when one processor is faulty because other processors can take over its scheduled tasks.

In this context, we investigate how to dynamically map and schedule tasks onto homogeneous faulty processors. We developed a run-time algorithm based on the primary/backup approach which is commonly used for its minimal resources utilization and high reliability. Its principal rule is that, when a task arrives, the system creates two identical copies: the primary copy and the backup copy. Several policies have been studied and their performances have been analyzed. We are currently refining the algorithm to reduce its complexity without decreasing performance. This work is done in collaboration with Oliver Sinnen, PARC Lab., the University of Auckland.

7.2.5. Energy Constrained and Real-Time Scheduling and Mapping on Multicores

Participants: Olivier Sentieys, Angeliki Kritikakou, Lei Mo.

Multicore architectures are now widely used in energy-constrained real-time systems, such as energy-harvesting wireless sensor networks. To take advantage of these multicores, there is a strong need to balance system energy, performance and Quality-of-Service (QoS). The Imprecise Computation (IC) model splits a task into mandatory and optional parts allowing to tradeoff QoS. We focus on the problem of mapping, i.e. allocating and scheduling, IC-tasks to a set of processors to maximize system QoS under real-time and energy constraints, which we formulate as a Mixed Integer Linear Programming (MILP) problem. However, state-of-the-art solving techniques either demand high complexity or can only achieve feasible (suboptimal) solutions. We develop an effective decomposition-based approach in [40] to achieve an optimal solution while reducing computational complexity. It decomposes the original problem into two smaller easier-to-solve problems: a master problem for IC-tasks allocation and a slave problem for IC-tasks scheduling. We also provide comprehensive optimality analysis for the proposed method. Through the simulations, we validate and demonstrate the performance of the proposed method, resulting in an average 55% QoS improvement with regards to published techniques.

7.2.6. Real-Time Scheduling of Reconfigurable Battery-Powered Multi-Core Platforms

Participants: Daniel Chillet, Aymen Gammoudi.

Reconfigurable real-time embedded systems are constantly increasingly used in applications like autonomous robots or sensor networks. Since they are powered by batteries, these systems have to be energy-aware, to adapt to their environment and to satisfy real-time constraints. For energy harvesting systems, regular recharges of battery can be estimated, and by including this parameter in the operating system, it is then possible to develop strategy able to ensure the best execution of the application until the next recharge. In this context, operating system services must control the execution of tasks to meet the application constraints. Our objective concerns the proposition of a new real-time scheduling strategy that considers execution constraints such as the deadline of tasks and the energy for heterogeneous architectures. For such systems, we first addressed homogeneous architectures and extended our work for heterogeneous systems for which each task has different execution parameters. For these two architectures models, we formulated the problem as an ILP optimisation problem that can be solved by classical solvers. Assuming that the energy consumed by the communication is dependent on the distance between processors, we proposed a mapping strategy to minimise the total cost of communication between processors by placing the dependent tasks as close as possible to each other. The proposed strategy guarantees that, when a task is mapped into the system and accepted, it is then correctly executed prior to the task deadline. Finally, as on-line scheduling is targeted for this work, we proposed heuristics to solve these problems in efficient way. These heuristics are based on the previous packing strategy developed for the mono-processor architecture case.

7.2.7. Run-Time Management on Multicore Platforms

Participant: Angeliki Kritikakou.

In real-time mixed-critical systems, Worst-Case Execution Time analysis (WCET) is required to guarantee that timing constraints are respected —at least for high criticality tasks. However, the WCET is pessimistic compared to the real execution time, especially for multicore platforms. As WCET computation considers the worst-case scenario, it means that whenever a high criticality task accesses a shared resource in multi-core platforms, it is considered that all cores use the same resource concurrently. This pessimism in WCET computation leads to a dramatic under utilization of the platform resources, or even failing to meet the timing constraints. In order to increase resource utilization while guaranteeing real-time guarantees for high criticality tasks, previous works proposed a run-time control system to monitor and decide when the interferences from low criticality tasks cannot be further tolerated. However, in the initial approaches, the points where the controller is executed were statically predefined. We propose a dynamic run-time control in [19] which adapts its observations to on-line temporal properties, increasing further the dynamism of the approach, and mitigating the unnecessary overhead implied by existing static approaches. Our dynamic adaptive approach allows to control the ongoing execution of tasks based on run-time information, and increases further the gains in terms of resource utilization compared with static approaches.

CELTIQUE Project-Team

4. New Results

4.1. Higher-Order Process Calculi

Participants: Sergueï Lenglet, Alan Schmitt.

Sergueï Lenglet and Alan Schmitt, in collaboration with researchers at Wrocław university, designed a fully abstract encoding of the λ-calculus into HOcore, a minimal higher-order process calculus. This work has been published at LICS [37]. In parallel, Lenglet and Schmitt have formalized HOπ in Coq and showed that its bisimilarity is compatible using Howe's method. This work has been accepted for publication at CPP 2018 [30].

4.2. Certified Semantics and Analyses for JavaScript

Participants: Gurvan Cabon, Alan Schmitt.

Alan Schmitt has continued his collaboration with Arthur Charguéraud (Inria Nancy) and Thomas Wood (Imperial College London) to develop JSExplain, an interpreter for JavaScript that is as close as possible to the specification. The tool is publicly available at https://github.com/jscert/jsexplain and is being extended to cover the current version of the standard.

In parallel, Gurvan Cabon and Alan Schmitt have developed a framework to automatically derive an information-flow tracking semantics from a pretty-big-step semantics. This work has been published [34] and is being formalized in Coq.

4.3. Certified Concurrent Garbage Collector

Participants: Yannick Zakowski, David Cachera, Delphine Demange, David Pichardie.

Concurrent garbage collection algorithms are an emblematic challenge in the area of concurrent program verification. We addressed this problem by proposing a mechanized proof methodology based on the popular Rely-Guarantee (RG) proof technique. We designed a specific compiler intermediate representation (IR) with strong type guarantees, dedicated support for abstract concurrent data structures, and high-level iterators on runtime internals (objects, roots, fields, thread identifiers...). In addition, we defined an RG program logic supporting an incremental proof methodology where annotations and invariants can be progressively enriched. We have formalized the IR, the proof system, and proved the soundness of the methodology in the Coq proof assistant. Equipped with this IR, we have proved the correctness of a fully concurrent garbage collector where mutators never have to wait for the collector. This work has been published in [32].

In this work, reasoning simultaneously about the garbage collection algorithm and the concrete implementation of the concurrent data-structures it uses would have entailed an undesired and unnecessary complexity. The above proof is therefore conducted with respect to abstract operations which execute atomically. In practice, however, concurrent data-structures uses fine-grained concurrency, for performance reasons. One must therefore prove an observational refinement between the abstract concurrent data-structures and their fined-grained, "linearisable" implementation. To adress this issue, we introduce a methodology inspired by the work of Vafeiadis, and provide the approach with solid semantic foundations. Assuming that fine-grained implementations are proved correct with respect to an RG specification encompassing linearization conditions, we prove, once and for all, that this entails a semantic refinement of their abstraction. This methodology is instantiated to prove correct the main data-structure used in our garbage collector. This work has been published in [33].

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

Participants: Thomas Genet, Thomas Jensen, Timothée Haudebourg.

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. Regular tree languages are (possibly) infinite languages which can be finitely represented using tree automata. To over-approximate sets of reachable terms, the tools we develop use the Tree Automata Completion (TAC) algorithm 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. We have already shown that tree automata completion can safely over-approximate the image of any first-order complete and terminating functional program. We have extended this result to the case of higher-order functional programs [40] and obtained very encouraging experimental results http://people.irisa.fr/Thomas.Genet/timbuk/funExperiments/. Besides, we have shown that completion was abble to take the evaluation strategy of the program into account [19]. The next step is to show the completeness of the approach, i.e., that any regular approximation of the image of a function can be found using completion. We already made progress in this direction [39].

4.5. C Semantics and Certified Compilation

Participants: Frédéric Besson, Sandrine Blazy.

The CompCert C compiler provides the formal guarantee that the observable behaviour of the compiled code improves on the observable behaviour of the source code. A first limitation of this guarantee is that if the source code goes wrong, i.e. does not have a well-defined behaviour, any compiled code is compliant. Another limitation is that CompCert 's notion of observable behaviour is restricted to IO events.

Over the past years, we have refined the semantics underlying CompCert so that (unlike CompCert but like Gcc) the binary representation of pointers can be manipulated much like integers and such that memory is a finite resource. We have now a formally verified C compiler, CompCertS, which is essentially the CompCert compiler, albeit with a stronger formal guarantee. The semantics preservation theorem applies to a wider class of existing C programs and, therefore, their compiled version benefits from the formal guarantee of CompCertS. CompCertS preserves not only the observable behaviour of programs but also ensures that the memory consumption is preserved by the compiler. As a result, we have the formal guarantee that the compiled code requires no more memory than the source code. This ensures that the absence of stack-overflows is preserved by compilation.

The whole proof of CompCertS represents a significant proof-effort. Details about the formal definition of the semantics and the proof of compiler passes can be found in the following publications [17], [25]

4.6. Constant-time verification by compilation and static analysis

Participants: Sandrine Blazy, David Pichardie, Alix Trieu.

To protect their implementations, cryptographers follow a very strict programming discipline called constant-time programming. They avoid branchings controlled by secret data as an attacker could use timing attacks, which are a broad class of side-channel attacks that measure different execution times of a program in order to infer some of its secret values. Several real-world secure C libraries such as NaCl, mbedTLS, or Open Quantum Safe, follow this discipline. We propose an advanced static analysis, based on state-of-the-art techniques from abstract interpretation, to report time leakage during programming. To that purpose, we analyze source C programs and use full context-sensitive and arithmetic-aware alias analyses to track the tainted flows. We give semantic evidences of the correctness of our approach on a core language. We also present a prototype implementation for C programs that is based on the CompCert compiler toolchain and its companion Verasco static analyzer. We present verification results on various real-world constant-time programs and report on a successful verification of a challenging SHA-256 implementation that was out of scope of previous tool-assisted approaches. This work has been published at ESORICS'17 [27].

The previous technique is well-adapted to verify the constant-time discipline at source level and give feedback to programmers, but the final security property must be established on the executable form of the program. In a joint work with IMDEA Software (Gilles Barthe and Vincent Laporte), we propose an automated methodology for validating on low-level intermediate representations the results of a source-level static analysis. Our methodology relies on two main ingredients: a relative-safety checker, an instance of a relational verifier which proves that a program is *safer* than another, and a transformation of programs into defensive form which verifies the analysis results at runtime. We prove the soundness of the methodology, and provide a formally verified instantiation based on the Verasco verified C static analyzer and the CompCert verified C compiler. This work has been published at CSF'17 [24].

CIDRE Project-Team

7. New Results

7.1. Intrusion Detection

7.1.1. Intrusion Detection in Distributed Systems

Alert Correlation: 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 the context of the PhD of Erwan Godefroy, 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. This is a top-down approach to correlation rule generation. With the PhD of Charles Xosanavongsa, we tackle the problem of a bottom-up approach that consists in discovering automatically the events or alerts that have been produced by the attacker activity. The objective is to classify automatically all suspicious entries in heterogeneous logs relative to a given attack. This requires to exhibit all log entries that are causally linked, and permits to produce a correlation rule that could detect later a new occurence of the attack.

Intrusion Detection in Cloud Infrastructure: Prior to detecting intrusion, it can be useful to know how the supervised system is vulnerable to attacks. Such result is obtained during a risk analysis phase in usual systems. In the PhD thesis of Pernelle Mensah, we try to automate the generation of the description of all possible attacks against a Cloud infrastructure. This work is divided in two separate steps: (1) We first discover the topology of the virtual machines executing in the cloud infrastructure [16], [17] and (2) Build in a second phase a topological attack graph that represents all possible known attacks on the virtual infrastructure. This graph will be later used either to adapt counter-measures to known attacks, or to generate automatically correlation rules to detect the described attacks.

Inferring the normal behavior of an application: We propose an 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 (machine learning approach). 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 current work solves this problem by characterizing the acceptable behaviors with invariant properties that they must verify. During the PhD thesis of David Lanoe, we enhanced the model building. Moreover, we assess this solution, by applying it to a distributed file system called XtreemFS. We show that it is possible to build the model of this given application, and to detect attack against XtreemFS, without producing too much false positives.

This approach is particularly appealing to detect intrusions in industrial control systems since these systems exhibit well-defined behaviors at different levels: network level (network communication patterns, protocol specifications, etc.), control level (continue and discrete process control laws), or even the state of the local resources (memory or CPU). Industrial control systems (ICS) can be subject to highly sophisticated attacks which may lead the process towards critical states. Due to the particular context of ICS, protection mechanisms

are not always practical, nor sufficient. On the other hand, developing a process-aware intrusion detection solution with satisfactory alert characterization remains an open problem. Sophisticated process-aware attacks targeting industrial control systems require adequate detection measures taking into account the physical process. We propose an approach relying on automatically mined process specifications to detect attacks on sequential control systems. The specifications are synthesized as monitors that read the execution traces and report violations to the operator. In contrast to other approaches, a central aspect of our method consists in reducing the number of mined specifications suffering from redundancies. We evaluate our approach on a hardware-in-the-loop testbed with a complex physical process model and discuss our approach's mining efficiency and attack detection capabilities. This work has been submitted to the Safeprocess'18 conference.

7.1.2. Illegal Information Flow Detection

Our research work on intrusion detection based on information flow has been initiated in 2002. This research work has resulted in Blare, a framework for Intrusion Detection Systems [0], including KBlare, an implementation as a Linux Security Module (LSM), JBlare, an implementation for the Java Virtual Machine (JVM), and AndroBlare, for Android applications.

Information Leaks: Qualitative information flow aims at detecting information leaks, whereas the emerging quantitative techniques target the estimation of information leaks. Quantifying information flow in the presence of low inputs is challenging, since the traditional techniques of approximating and counting the reachable states of a program no longer suffice. We propose an automated quantitative information flow analysis for imperative deterministic programs with low inputs. The approach relies on a novel abstract domain, the cardinal abstraction, in order to compute a precise upper-bound over the maximum leakage of batch-job programs. We prove the soundness of the cardinal abstract domain by relying on the framework of abstract interpretation. We also prove its precision with respect to a flow-sensitive type system for the two-point security lattice. This approach has been published in POPL'17 [8].

Correct information flow monitoring by design: As mentionned previously, our research team is developing an information monitor called Blare. Like most of its competitors (e.g. Laminar or Weir) our solution is based on the Linux Security Module (LSM) framework. However, this framework was initially designed with access control in mind. A natural question arises from this matter of fact: does the LSM framework can be used to correctly track information flow (at the operating system level) ? In the context of his PhD thesis, Laurent Georget has studied this very same question.

To tackle this problem, Laurent Georget has designed an ad hoc static analysis that run as a GCC plugin during the Linux kernel compilation. This analysis can prove (or disprove) the fact that LSM hooks within a chosen set of system calls (known to realize information flows between operating systems containers like files, sockets or pipe) are placed at correct locations so as to intercept these possible information flows. The experiments conducted by Laurent Georget have revealed that on an initial set of 38 system calls, 28 were correctly instrumented by LSM, 4 of them were equipped with a LSM hook that could miss some information flow (under certain circumstances), 3 were simply lacking a LSM hook, and 3 false positives had to be manually analyzed and requalified. Laurent Georget was able to produce a kernel patch to remove all missing and misplaced hooks. This patch can be prove to be correct using the same tool. This contribution was published at FormaliSE 2017 [12].

We had detected for a long time a subtle bug in our information flow monitor implementation (Blare) that we were able to track down to a race condition between two concurrent system calls reading and writing into the same pipe. Laurent Georget has proposed during its PhD an elegant solution to this complex problem: he proposed to divide each information flow into three stages: the activation, the execution and the deactivation. Only the activation and deactivation can be observed by the monitor using LSM hooks placed at the beginning and the exit of a system call. This way, it becomes possible to track causal dependencies between concurrent system calls within the LSM framework. Laurent Georget has proved (using the Coq proof assistant) that his approach is correct and computes the smallest possible over-approximation, in the sense that for any concurrent execution where multiple system calls are used there exists a linearization of this execution that produces the

[0] http://www.blare-ids.org/

information flow computed by his algorithm. Laurent Georget has implemented his algorithm in the Linux kernel. This contribution was publish at Software Engineering & Formal Methods (2017) where it was granted the best paper award [11]. Laurent Georget has defended his PhD thesis in September 2017 .

Advanced Persistent Threats: Long lived attack campaigns known as Advanced Persistent Threats (APTs) have emerged as a serious security risk. These attack campaigns are customised for their target and performed step by step during months on end. The major difficulty in detecting an APT is keeping track of the different steps logged over months of monitoring and linking them. In [29], we described TerminAPTor, an APT detector which highlights links between the traces left by attackers in the monitored system during the different stages of an attack campaign. TerminAPTor tackles this challenge by resorting to Information Flow Tracking (IFT). TerminAPTor was presented last year and we have pursue our effort in this area. More precisely, we have focus on the evaluation of this solution and thus we face to the lack of public datasets of attacks. We develop Moirai a framework dedicated to attacks scenario sharing [22] .

Characterizing Android Malware: Android has become the world's most popular mobile operating system, and consequently the most popular target for unscrupulous developers. These developers seek to make money by taking advantage of Android users who customize their devices with various applications, which are the main malware infection vector. Indeed, the most likely way a user executes a repackaged application is by downloading a seemingly harmless application from a store and executing it. Such an application may have been modified by an attacker in order to add malicious pieces of code.

To fight repackaged applications containing malicious code, most official application marketplaces have implemented security analysis tools that try to detect and remove malware. Countermeasures adopted by the attackers to bypass these new controls can be divided into two main approaches: avoiding static analysis and avoiding dynamic analysis. A static analysis of an application consists of analysing its code and its resources without executing it. Conversely, dynamic analysis stands for any kind of analysis that requires executing the application in order to observe its actions.

The Kharon project [30] goes a step further from classical dynamic analysis of malware [0]. Funded by the Labex CominLabs and involving partners of CentraleSupélec, Inria and INSA Centre Val de Loire, this project aims to capture a compact and comprehensive representation of malware. To achieve such a goal we have developed tools to monitor operating systems' information flows induced by the execution of a marked application. We support the idea that the best way to understand malware impact is to observe it in its normal execution environment i.e., a real smartphone. Additionally, the main challenge is to be able to trigger malicious behaviors even when the malware tries to escape dynamic analysis.

In this context, we have developed an original solution whose main purpose is a relavant dynamic analysis of the malicious code. We develop the GroddDroid software, that mainly consists of 'helping the malware to execute'. To reach this goal, GroddDroid relies on a previous static analysis that evidence all the execution paths leading to the malicious code. We compute a global control flow graph (CFG) that exhibits execution paths to reach specific parts of code, even if these paths use callbacks that are handled in the Android framework itself [15]. Finally, GroddDroid slightly modifies the bytecode of the infected application in order to defeat the protection against dynamic analysis and executes the suspicious code in its most favorable execution conditions. Thus, GroddDroid helps to understanding malware's objectives and the consequences on the health of a user's device.

GroddDroid can also be used for classifying applications between goodware and malware. We show in [19] that benign applications have a System Flow Graph (a graph that represent flows at operating system level) that can be anticipated. Malware that perform complex operations such as installing backdoor or launching a Tor client, have a CFG that differ enough to be classified easily.

Our main research direction and challenges in this area are to continue to enhance these technologies in order to reach a sufficient level of software maturity to deploy a permanent platform of malware analysis in the LHS (Laboratory of High Security) and to create new opportunities with industrial partners.

[0] http://kharon.gforge.inria.fr

7.1.3. Intrusion Detection in Low-Level Software Components

In order to protect the IDS itself, we have initiated different research activities in the domain of hardware security. Our goal is to use co-design software/hardware approaches against traditional software attacks. In a bilateral research project with HP Inc Research Labs, we investigate how dedicated hardware could be used to monitor the whole software stack (from the firmware to the user-mode applications). In the CominLabs HardBlare project, we study the use of a dedicated co-processor to enforce Information Flow Control (IFC) on the main CPU. Finally, in the context of the PhD thesis of Thomas Letan (ANSSI), we investigate the use of formal methods to evaluate the security guarantees provided by hardware platforms, which combine different CPUs, chipsets and memories.

Highly privileged software, such as firmware, is an attractive target for an attacker. Thus, BIOS vendors use cryptographic signatures to ensure firmware integrity at boot time. Nevertheless, such boot time protection does not prevent an attacker from exploiting vulnerabilities at runtime. To detect such runtime attacks, we proposed an event-based monitoring approach that relies on an isolated co-processor [10]. We instrument the code executed on the main CPU to send information about its behavior to the monitor. In this work, we focus on the detection of attacks targeting the System Management Mode (SMM), a highly privileged x86 execution mode executing firmware code at runtime. We use the control flow of the code as a model of its behavior. We evaluate our approach with two open-source implementations: EDK II and coreboot. We evaluate its ability to detect state-of-the-art attacks and its runtime execution overhead by simulating an x86 system coupled with an ARM Cortex A5 co-processor. The results show that our solution detects intrusions from the state of the art while remaining acceptable in terms of performance overhead in the context of the SMM. This work has been done in collaboration with HP Inc Research Labs, in the context of the PhD of Ronny Chevalier.

Over time, hardware designs have constantly grown in complexity and modern platforms involve multiple interconnected hardware components. During the last decade, several vulnerability disclosures have proven that trust in hardware can be misplaced. The approach we developed with Thomas Letan rely on a formal definition of Hardware-based Security Enforcement (HSE) mechanisms, a class of security enforcement mechanisms such that a software component relies on the underlying hardware platform to enforce a security policy. We then model a subset of a x86-based hardware platform specifications and we prove the soundness of a realistic HSE mechanism within this model using Coq, a proof assistant system.

The HardBlare project proposes a software/hardware co-design methodology to ensure that security properties are preserved all along the execution of the system but also during files storage. It is based on the Dynamic Information Flow Tracking (DIFT) that generally consists in attaching tags to denote the type of information that are saved or generated within the system. These tags are then propagated when the system evolves and information flow control is performed in order to guarantee the safe execution and storage within the system monitored by security policies. We proposed ARMHEx [20], a practical solution targeting DIFT on ARM-based SoCs (e.g. Xilinx Zynq). Current DIFT implementations suffer from two major drawbacks. First, recovering required information for DIFT is generally based on software instrumentation leading to high time overheads. ARMHEx takes profit of ARM CoreSight debug components and static analysis to drastically reduce instrumentation time overhead (up to 90% compared to existing works). Then, security of the DIFT hardware extension itself is not considered in related works. In this work, we tackle this issue by proposing a solution based on ARM Trustzone. This work has been done in the context of the PhD of Muhammad Abdul Wahab and Mounir Nasr Allah.

7.1.4. Vizualization

When using Intrusion Detection Systems (IDS), the large quantities of alerts generated are difficult to handle by security experts. To help solving this problem, we have proposed VEGAS, an alerts visualization and classification tool that allows primary visions based on their principal component analysis (PCA) representation. Following this, we have studied the context of collaboration between the various security actors. We have then proposed an extension to VEGAS that allows to help the actors to collaborate. We have developped an interface that permits the front-end operator to quickly understand the security events, and group them to organize incidents and send them to dedicated analysts. Conversely, once the incidents have

been analysed, the analysts can send information to the front-line operators to help them understanding the futur security events.

We also developed another tool called STARLORD [14] that permits to an administrator the explore in a 3D graph representing the links between the heterogeneous entries in various logs produced either by the system, applications or IDSes. To emphasize the important relations between the lines of logs that can potentially be part of an attack activity, we classify these links in order to present only the part of the graph that is linked to an indicators of compromission.

Our previous research on visualization of security events has lead to two proofs-of-concept (See ELVIS and CORGI softwares). We are currently pursuing business opportunities on this topic. Indeed SplitSec is a soon to be founded startup developing tools to help security experts to better manage and understand security data. Scalable analysis solutions and data visualisations adapted for security are combined into powerful tools for incident response. Until June 2017, Christopher Humphries has been hired by Inria as a technology transfer engineer to build these tools based on promising research prototypes.

7.2. Privacy

7.2.1. Image Encryption

More and more users prefer to share their photos through image-sharing platforms of social networks than using e-mail or personal webpages. Since the provider of the image-sharing platform can clearly know the contents of any published images, the users have to trust the provider to respect their privacy or has to encrypt their images. In the context of the PhD of Kun He, we have proposed an IND-CPA image encryption algorithm that preserve the image format after encryption, and we have shown that our encryption algorithm can be used on several widely used image-sharing platforms such as Flickr, Pinterest, Google+ and Twitter. Kun He has completed her PhD thesis in September 2017 [5].

7.3. Security of Communicating and Distributed Systems

7.3.1. Routing Protocol for Tactical Mobile Ad Hoc Networks

In the context of the PhD thesis of Florian Grandhomme, we propose new secure and efficient algorithms and protocols to provide inter-domain routing in the context of tactical mobile ad hoc network. The proposed protocol has to handle context modification due to the mobility of Mobile Ad hoc NETwork (MANET), that is to say split of a MANET, merge of two or more MANET, and also handle heterogeneity of technology and infrastructure. The solution has to be independent from the underlying intra-domain routing protocol and from the infrastructure: wired or wireless, fixed or mobile. This work is done in cooperation with DGA-MI.

New generation military equipment, soldiers and vehicles, use wireless technology to communicate on the battlefield. During missions, they form a MANET. Since the battlefield includes coalition, each group may communicate with another group, and inter-MANET communication may be established. Inter-MANET (or inter-domain MANET) communication should allow communication, but maintain a control on the exchanged information. Several protocols have been proposed in order to handle inter-domain routing for tactical MANETs. During the thesis we have shown that simulator (NS3) or emulator (CORE) do not handle correctly ad hoc network behavior and then that solution in the state of the art are more complex than needed. Based on this analysis, we propose some preconizations to design Inter-domain protocols for MANET and we propose the ITMAN (Inter Tactical Mobile Ad hoc Network) protocol that allows also to handle simple routing policy (merge, link and deny). We evaluate this new protocol through experimentation and we show that our proposition is quite efficient. On going work on this protocol is the definition and implementation of more subtle routing policy that allow announce filtering of giving prefix for example.

7.3.2. Decentralized Cryptocurrency Systems

Distributed Ledgers (e.g. Bitcoin) occupy currently the first lines of the economical and political media and many speculations are done with respect to their level of coherence and their computability power. Interestingly, there is no consensus on the properties and abstractions that fully capture the behaviour of distributed ledgers. The interest in formalising the behaviour of distributed ledgers is twofold. Firstly, it helps to prove the correctness of the algorithms that implement existing distributed ledgers and explore their limits with respect to an unfriendly environment and target applications. Secondly, it facilitates the identification of the minimal building blocks necessary to implement the distributed ledger in a specific environment. Even though the behaviour of distributed ledgers is similar to abstractions that have been deeply studied for decades in distributed systems no abstraction is sufficiently powerful to capture the distributed ledger behaviour. We have defined the Distributed Ledger Register, a register that mimics the behaviour of one of the most popular distributed ledger, i.e. the Bitcoin ledger. The aim of our work is to provide formal guarantees on the coherent evolution of Bitcoin. We furthermore showed that the Bitcoin blockchain maintenance algorithm verifies the distributed ledger register properties under strict conditions. Moreover, we proved that the Distributed Ledger Register verifies the regularity register specification. It follows that the strongest coherency implemented by Bitcoin is regularity under strong assumptions (i.e. partial synchronous systems and sparse reads). In [7] we proposed a study that contradicts the common belief that Bitcoin implements strong coherency criteria in a totally asynchronous system. To the best of our knowledge, our work is the first one that makes the connection between the distributed ledgers and the classical theory of distributed shared registers.

Double spending and blockchain forks are two main issues that the Bitcoin crypto-system is confronted with. The former refers to an adversary's ability to use the very same coin more than once while the latter reflects the occurrence of transient inconsistencies in the history of the blockchain distributed data structure. We present a new approach to tackle these issues: it consists in adding some local synchronization constraints on Bitcoin's validation operations, and in making these constraints independent from the native blockchain protocol. Synchronization constraints are handled by nodes which are randomly and dynamically chosen in the Bitcoin system. In [13] we show that with such an approach, content of the blockchain is consistent with all validated transactions and blocks which guarantees the absence of both double-spending attacks and blockchain forks.

7.3.3. Large Scale Systems

Population Protocol: the computational model of population protocols is a formalism that allows the analysis of properties emerging from simple and pairwise interactions among a very large number of anonymous finite-state agents. Significant work has been done so far to determine which problems are solvable in this model and at which cost in terms of states used by the protocols and time needed to converge. The problem tackled in is the population proportion problem: each agent starts independently from each other in one of two states, say A or B, and the objective is for each agent to determine the proportion of agents that initially started in state A, assuming that each agent only uses a finite set of state, and does not know the number n of agents. In [18], we show that for any $\delta \in (0,1)$, the number of interactions needed per node to converge is $O(ln(n/\delta))$ with probability at least $1-\delta$. We also prove that each node can determine, with any high probability, the proportion of nodes that initially started in a given state without knowing the number of nodes in the system. This work provides a precise analysis of the convergence bounds, and shows that using the 4-norm is very effective to derive useful bounds.

Distributed Stream Processing Systems: shuffle grouping is a technique used by stream processing frameworks to share input load among parallel instances of stateless operators. With shuffle grouping each tuple of a stream can be assigned to any available operator instance, independently from any previous assignment. A common approach to implement shuffle grouping is to adopt a Round-Robin policy, a simple solution that fares well as long as the tuple execution time is almost the same for all the tuples. However, such an assumption rarely holds in real cases where execution time strongly depends on tuple content. As a consequence, parallel stateless operators within stream processing applications may experience unpredictable unbalance that, in the end, causes undesirable increase in tuple completion times. In [25] we propose Online Shuffle Grouping

(OSG), a novel approach to shuffle grouping aimed at reducing the overall tuple completion time. OSG estimates the execution time of each tuple, enabling a proactive and online scheduling of input load to the target operator instances. Sketches are used to efficiently store the otherwise large amount of information required to schedule incoming load. We provide a probabilistic analysis and illustrate, through both simulations and a running prototype, its impact on stream processing applications.

The real time analysis of massive data streams is of utmost importance in data intensive applications that need to detect as fast as possible and as efficiently as possible (in terms of computation and memory space) any correlation between its inputs or any deviance from some expected nominal behavior. The IoT infrastructure can be used for monitoring any events or changes in structural conditions that can compromise safety and increase risk. It is thus a recurrent and crucial issue to determine whether huge data streams, received at monitored devices, are correlated or not as it may reveal the presence of attacks. We propose a metric, called codeviation, that allows to evaluate the correlation between distributed massive streams. This metric is inspired from classical metric in statistics and probability theory, and as such enables to understand how observed quantities change together, and in which proportion. In [6], we 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 then generalize our approach by presenting a new metric, the Sketch-metric, which allows us to define a distance between updatable summaries of large data streams. An important feature of the Sketch-metric is that, given a measure on the entire initial data streams, the Sketch-metric preserves the axioms of the latter measure on the sketch. We finally conducted extensive experiments on both synthetic traces and real data sets allowing us to validate the robustness and accuracy of our metrics.

HYCOMES Project-Team

5. New Results

5.1. Semantics, Static or Runtime Analysis of Hybrid Systems

5.1.1. Structural Analysis of Multi-Mode DAEs

Differential Algebraic Equation (DAE) systems constitute the mathematical model supporting physical modeling languages such as Modelica or Simscape. Unlike Ordinary Differential Equations, or ODEs, they exhibit subtle issues because of their implicit *latent equations* and related *differentiation index*. Multi-mode DAE (mDAE) systems are much harder to deal with, not only because of their mode-dependent dynamics, but essentially because of the events and resets occurring at mode transitions. Unfortunately, the large literature devoted to the numerical analysis of DAEs do not cover the multi-mode case. It typically says nothing about mode changes. This lack of foundations cause numerous difficulties to the existing modeling tools. Some models are well handled, others are not, with no clear boundary between the two classes. In [11], we develop a comprehensive mathematical approach to the *structural analysis* of mDAE systems which properly extends the usual analysis of DAE systems. We define a constructive semantics based on nonstandard analysis and show how to produce execution schemes in a systematic way. This work has been accepted for presentation at the HSCC 2017 conference [18] in April 2017.

5.1.2. Operational Models for Piecewise-Smooth Systems

In [7], we study ways of constructing meaningful operational models of piecewise-smooth systems (PWS). The systems we consider are described by polynomial vector fields defined on non-overlapping semi-algebraic sets, which form a partition of the state space. Our approach is to give meaning to motion in systems of this type by automatically synthesizing operational models in the form of hybrid automata (HA). Despite appearances, it is in practice often difficult to arrive at satisfactory HA models of PWS. The different ways of building operational models that we explore in our approach can be thought of as defining different semantics for the underlying PWS. These differences have a number of interesting nuances related to phenomena such as chattering, non-determinism, so-called mythical modes and sliding behaviour.

5.1.3. Accelerated Simulation of Hybrid Systems: Method combining static analysis and runtime execution analysis

Ayman Aljarbouh has defended his PhD [4] on September 13th 2017. His PhD has been partially funded by an ARED grant of the Brittany Regional Council. His doctoral work took place in the context of the Modrio (completed in 2016) and Sys2Soft (completed in 2015) projects on hybrid systems modeling. Ayman Aljarbouh has been working on accelerated simulation techniques for hybrid systems. In particular, he has contributed, and implemented in a software prototype, a regularisation method transforming automatically at runtime a chattering behaviour into a semantics preserving smooth behaviour. He has also contributed a method for the approximation of Zeno behaviour. This method enables to jump past an accumulation of an infinite number of zero-crossing events, and to continue the simulation of a large class of Zeno hybrid systems, after accumulation points.

5.1.4. A Type-based Analysis of Causality Loops in Hybrid Systems Modelers

Explicit hybrid systems modelers like Simulink/Stateflow allow for programming both discrete- and continuous-time behaviors with complex interactions between them. A key issue in their compilation is the static detection of algebraic or causality loops. Such loops can cause simulations to deadlock and prevent the generation of statically scheduled code. In [5], we addresses this issue for a hybrid modeling language that combines synchronous data-flow equations with Ordinary Differential Equations (ODEs). We introduce the operator last(x) for the left-limit of a signal x. This operator is used to break causality loops and permits a

uniform treatment of discrete and continuous state variables. The semantics relies on non-standard analysis, defining an execution as a sequence of infinitesimally small steps. A signal is deemed causally correct when it can be computed sequentially and only changes infinitesimally outside of announced discrete events like zero-crossings. The causality analysis takes the form of a type system that expresses dependences between signals. In well-typed programs, signals are provably continuous during integration provided that imported external functions are also continuous. The effectiveness of this system is illustrated with several examples written in Zélus, a Lustre-like synchronous language extended with hierarchical automata and ODEs.

5.2. Formal Verification of Hybrid Systems

5.2.1. Formal Verification of Station Keeping Maneuvers for a Planar Autonomous Hybrid System

In [9], we investigate the formal verification of a hybrid control law designed to perform a station keeping maneuver for a planar vehicle. Such maneuver requires that the vehicle reaches a neighborhood of its station in finite time and remains in it while waiting for further commands. We model the dynamics as well as the control law as a hybrid program and formally verify the reachability and safety properties involved. We highlight in particular the automated generation of invariant regions which turns out to be crucial in performing such verification. We use the hybrid system theorem prover KeymaeraX to formally check the parts of the proof that can be automatized in the current state of the tool.

5.2.2. Formal verification of obstacle avoidance and navigation of ground robots

In [6], we answer fundamental safety questions for ground robot navigation: Under which circumstances does a given control decision make a ground robot safely avoid obstacles? Unsurprisingly, the answer depends on the exact formulation of the safety objective as well as the physical capabilities and limitations of the robot and the obstacles. Because uncertainties about the exact future behavior of a robot's environment make this a challenging problem, we formally verify corresponding controllers and provide rigorous safety proofs justifying why they can never collide with the obstacle in the respective physical model. To account for ground robots in which different physical phenomena are important, we analyze a series of increasingly strong properties of controllers for increasingly rich dynamics and identify the impact that the additional model parameters have on the required safety margins. We analyze and formally verify: (i) static safety, which ensures that no collisions can happen with stationary obstacles, (ii) passive safety, which ensures that no collisions can happen with stationary or moving obstacles while the robot moves, (iii) the stronger passive friendly safety in which the robot further maintains sufficient maneuvering distance for obstacles to avoid collision as well, and (iv) passive orientation safety, which allows for imperfect sensor coverage of the robot, i. e., the robot is aware that not everything in its environment will be visible. We formally prove that safety can be guaranteed despite sensor uncertainty and actuator perturbation. We complement these provably correct safety properties with liveness properties: we prove that provably safe motion is flexible enough to let the robot navigate waypoints and pass intersections. In order to account for the mixed influence of discrete control decisions and the continuous physical motion of the ground robot, we develop corresponding hybrid system models and use differential dynamic logic theorem proving techniques to formally verify their correctness. Since these models identify a broad range of conditions under which control decisions are provably safe, our results apply to any control algorithm for ground robots with the same dynamics. As a demonstration, we, thus, also synthesize provably correct runtime monitor conditions that check the compliance of any control algorithm with the verified control decisions.

5.3. Synchronous Interfaces and Assume/Guarantee Contracts

In [10], we establish a link between the theory of Moore Interfaces proposed in 2002 by Chakraborty et al. as a specification framework for synchronous transition systems, and the Assume/Guarantee contracts as proposed in 2007 by Benveniste et al. as a simple and flexible contract framework. As our main result we show that the operation of saturation of A/G contracts (namely the mapping $(A, G) \rightarrow (A, G \vee \neg A)$), which was considered a drawback of this theory, is indeed implemented by the Moore Game of Chakraborty et al. We further develop this link and come up with some remarks on Moore Interfaces.

5.4. CominWeb project of the Labex CominLabs

Jean Hany and Albert Benveniste (together with William Dedzoe) were involved in this project.

CominWeb is a project supported by the Labex CominLabs since 2013. Its original objective was to equip CominLabs with Web infrastructures, tools, and services, that would allow to run the scientific activity of the Labex in an innovative way. Based on a study of the population of the CominLabs researchers, performed in year 2014-15 by the teams of CominLabs involved in social sciences, several services were investigated and prototyped. A short trial addressed the automatic generation of a scientific activity report, for a CominLabs project, from the material available from the publications ot the project team. This was suspended because such a service was not considered very useful by the community. A second trial (nicknamed "NSA") consisted in monitoring the flows of email exchanges addressed to aliases of the CominLabs projects, with the objective of classifying the mails into: meeting announcements, mails with attachments of interest, and other mails. This would give to the CominLabs head a view on the project's activities without asking for any specific contribution from the researchers. This was more interesting. Still, a difficulty was that researchers did not use the project aliases so much. For priority issues, this development was also suspended.

The main result of this project is thus the service called *LookinLabs*, deployed in two different versions: http://lookinlabs4halinria.cominlabs.ueb.eu/ and http://www.lookinlabs.cominlabs.ueb.eu/. The former is a more advanced version of LookinLabs, developed for the whole Inria community, by exploiting the HAL publication archive. LookinLabs for HAL-Inria allows the user to find, among teams/individuals/publications taken from all the Inria teams, those best matching a query consisting of a list of keywords or a short text. The tool exploits, as data, HAL-Inria archives, in combination with the Inria Activity reports (the Raweb), and the internal data base of Inria teams called BASTRI. Active teams/individuals are shown in boldface. Teams/individuals shown in gray are no longer active at Inria. If team TEAM0 is no longer active, the mention: TEAM0 $\rightarrow$ (TEAM1,TEAM2) indicates follow-up active teams, if any. In LookinLabs, no ontology is used. No data need to be manually entered (besides the users' queries). The tool uses *Elasticsearch* (https://www.elastic.co/fr/products/elasticsearch) as its core algorithm. This means that the matching is based on a distance between the query and the set of data attached, in HAL, to each team/individual/publication. Ranking is performed accordingly. Explanations are given for each returned item. Correlation graphs are given, allowing to navigate through teams or individuals that share common interests (they may or may not be co-authors).

LookinLabs is deployed in two versions. LookinLabs4HALInria is the one we just described. The other version is in operation since 2016 and addresses the scientific community of CominLabs researchers. The data used are up to 10 standard bibliographical data bases (Dblp, IEEE Explore, Arxiv, HAL, and more) for which links have been collected from the researchers (this was the only data they were asked for). Results are returned in the form of individuals and publications, not teams.

PACAP Project-Team

6. New Results

6.1. Compiler, vectorization, interpretation

Participants: Erven Rohou, André Seznec, Sylvain Collange, Rabab Bouziane, Arif Ali Ana-Pparakkal, Stefano Cherubin, Byron Hawkins, Arif Ali Ana-Pparakkal, Imane Lasri, Kévin Le Bon.

6.1.1. Improving sequential performance through memoization

Participants: Erven Rohou, Imane Lasri, André Seznec.

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 previously proposed [23] 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.

A restriction of the proposed framework was that memoization was restricted only to dynamically linked functions and the functions must be determined beforehand. We extended this work, and we propose function memoization using a compile-time technique thus extending the scope of memoization to user defined functions as well as making it transparently applicable to any dynamically linked functions. Our compile-time technique allows static linking of memoization code and this increases the benefit due to memoization by leveraging the inlining capability for the memoization wrapper. Our compile-time analysis can also handle functions with pointer parameters , and we handle constants more efficiently. Instruction set support can also be considered, and we propose associated hardware leading to additional performance gain.

This work was presented at the Compiler Construction Conference 2017 [50]. It is also described in the PhD thesis of Arjun Suresh [24].

6.1.2. Optimization in the Presence of NVRAM

Participants: Erven Rohou, Rabab Bouziane.

Beyond the fact of generating machine code, compilers play a critical role in delivering high performance, and more recently high energy efficiency. For decades, the memory technology of target systems has consisted in SRAM at cache level, and DRAM for main memory. Emerging non-volatile memories (NVMs) open up new opportunities, along with new design challenges. In particular, the asymmetric cost of read/write accesses calls for adjusting existing techniques in order to efficiently exploit NVMs. In addition, this technology makes it possible to design memories with cheaper accesses at the cost of lower data retention times. These features can be exploited at compile time to derive better data mappings according to the application and data retention characteristics. We reviewed a number of compile-time analysis and optimization techniques, and how they could apply to systems in presence of NVMs [37]. In particular, we consider the case of the reduction of the number of writes, and the analysis of variables lifetime for memory bank assignment of program variables.

Concerning the reduction of writes, we propose a fast evaluation of NVM integration at cache level, together with a compile-time approach for mitigating the penalty incurred by the high write latency of STT-RAM. We implement a code optimization in LLVM for reducing so-called *silent stores*, i.e., store instruction instances that write to memory values that were already present there. This makes our optimization portable over any architecture supporting LLVM. Then, we assess the possible benefit of such an optimization on the Rodinia benchmark suite through an analytic approach based on parameters extracted from the literature devoted to NVMs. This makes it possible to rapidly analyze the impact of NVMs on memory energy consumption. Reported results show up to 42 % energy gain when considering STT-RAM caches. This work is accepted for publication at RAPIDO'18 [38].

This research is done in collaboration with Abdoulaye Gamatié at LIRMM (Montpellier) within the context the the ANR project CONTINUUM.

6.1.3. Dynamic Binary Optimization

Participants: Erven Rohou, Arif Ali Ana-Pparakkal, Kévin Le Bon, Byron Hawkins.

6.1.3.1. Dynamic Function Specialization

Participants: Erven Rohou, Arif Ali Ana-Pparakkal, Kévin Le Bon.

Compilers can do better optimization with the knowledge of run-time behavior of the program. *Function specialization* is a compilation technique that consists in optimizing the body of a function for specific values of an argument. Different versions of a function are created to deal with the most frequent values of the arguments, as well as the default case. Compilers can do a better optimization with the knowledge of run-time behaviour of the program. Static compilers, however, can hardly predict the exact value/behaviour of arguments, and even profiling collected during previous runs is never guaranteed to capture future behaviour. We propose a dynamic function specialization technique, that captures the actual values of arguments during execution of the program and, when profitable, creates specialized versions and include them at runtime. Our approach relies on dynamic binary rewriting. We present [36] the principles and implementation details of our technique, analyze sources of overhead, and present our results.

This research is done within the context of the Nano 2017 PSAIC collaborative project.

6.1.3.2. Runtime Vectorization of Binary Programs

Participant: Erven Rohou.

In many cases, applications are not optimized for the hardware on which they run. Several reasons contribute to this unsatisfying situation, such as legacy code, commercial code distributed in binary form, or deployment on compute farms. In fact, backward compatibility of ISA guarantees only the functionality, not the best exploitation of the hardware. In this work, we focus on maximizing the CPU efficiency for the SIMD extensions.

We previously proposed [3] a binary-to-binary optimization framework where loops vectorized for an older version of the processor SIMD extension are automatically converted to a newer one. It is a lightweight mechanism that does not include a vectorizer, but instead leverages what a static vectorizer previously did. We showed that many loops compiled for x86 SSE can be dynamically converted to the more recent and more powerful AVX; as well as, how correctness is maintained with regards to challenges such as data dependencies and reductions. We obtained speedups in line with those of a native compiler targeting AVX.

We now focus on runtime vectorization of loops in binary codes that were not originally vectorized [29]. For this purpose, we use open source frameworks that we have tuned and integrated to

1. dynamically lift the x86 binary into the Intermediate Representation form of the LLVM compiler,
2. abstract hot loops in the polyhedral model,
3. use the power of this mathematical framework to vectorize them,
4. and finally compile them back into executable form using the LLVM Just-In-Time compiler.

In most cases, the obtained speedups are close to the number of elements that can be simultaneously processed by the SIMD unit. The re-vectorizer and auto-vectorizer are implemented inside a dynamic optimization platform; it is completely transparent to the user, does not require any rewriting of the binaries, and operates during program execution.

This work is done in collaboration with Philippe Clauss (Inria CAMUS), it is part of the PhD work of Nabil Hallou [26].

6.1.4. Hardware/Software JIT Compiler

Participant: Erven Rohou.

Dynamic Binary Translation (DBT) is often used in hardware/software co-design to take advantage of an architecture model while using binaries from another one. The co-development of the DBT engine and of the execution architecture leads to architecture with special support to these mechanisms. We proposed [46] a hardware accelerated dynamic binary translation where the first steps of the DBT process are fully accelerated in hardware. Results showed that using our hardware accelerators leads to a speed-up of 8× and a cost in energy 18× lower, compared with an equivalent software approach.

Single ISA-Heterogeneous multi-cores such as the ARM big.LITTLE have proven to be an attractive solution to explore different energy/performance trade-offs. Such architectures combine Out of Order cores with smaller in-order ones to offer different power/energy profiles. They however do not really exploit the characteristics of workloads (compute-intensive vs. control dominated). In our recent work, we propose to enrich these architectures with runtime configurable VLIW cores, which are very efficient at compute-intensive kernels. To preserve the single ISA programming model, we resort to Dynamic Binary Translation, and use this technique to enable dynamic code specialization for Runtime Reconfigurable VLIWs cores. Our proposed DBT framework targets the RISC-V ISA, for which both OoO and in-order implementations exist. Our experimental results show that our approach can lead to best-case performance and energy efficiency when compared against static VLIW configurations.

This work has been accepted for publication at DATE 2018 [53].

This research is done in collaboration with Steven Derrien and Simon Rokicki from the CAIRN team.

6.1.5. Customized Precision Computing

Participants: Erven Rohou, Stefano Cherubin, Imane Lasri.

Error-tolerating applications are increasingly common in the emerging field of real-time HPC. Proposals have been made at the hardware level to take advantage of inherent perceptual limitations, redundant data, or reduced precision input, as well as to reduce system costs or improve power efficiency. At the same time, works on floating-point to fixed-point conversion tools allow us to trade-off the algorithm exactness for a more efficient implementation. In this work [39], we aim at leveraging existing, HPC-oriented hardware architectures, while including in the precision tuning an adaptive selection of floating-and fixed-point arithmetic. Our proposed solution takes advantage of the application domain knowledge of the programmers by involving them in the first step of the interaction chain. We rely on annotations written by the programmer on the input file to know which variables of a computational kernel should be converted to fixed-point. The second stage replaces the floating-point variables in the kernel with fixed-point equivalents. It also adds to the original source code the utility functions to perform data type conversions from floating-point to fixed-point, and vice versa. The output of the second stage is a new version of the kernel source code which exploits fixed-point computation instead of floating-point computation. As opposed to typical custom-width hardware designs, we only rely on the standard 16-bit, 32-bit and 64-bit types. We also explore the impact of the fixed-point representation on auto-vectorization. We discuss the effect of our solution in terms of time-to-solutions, error and energy-to-solution.

This is done within the context of the ANTAREX project in collaboration with Stefano Cherubin, and Giovanni Agosta from Politecnico di Milano, and Olivier Sentieys from the CAIRN team.

6.1.6. SPMD Function Call Re-Vectorization

Participant: Sylvain Collange.

SPMD programming languages for SIMD hardware such as C for CUDA, OpenCL or ISPC have contributed to increase the programmability of SIMD accelerators and graphics processing units. However, SPMD languages still lack the flexibility offered by low-level SIMD programming on explicit vectors. To close this expressiveness gap while preserving the SPMD abstraction, we introduce the notion of Function Call Re-Vectorization (CREV). CREV allows changing the dimension of vectorization during the execution of an SPMD kernel, and exposes it as a nested parallel kernel call. CREV affords a programmability close to dynamic parallelism, a feature that allows the invocation of kernels from inside kernels, but at much lower cost. We defined a formal semantics of CREV, and implemented it on the ISPC compiler. To validate our

idea, we have used CREV to implement some classic algorithms, including string matching, depth first search and Bellman-Ford, with minimum effort. These algorithms, once compiled by ISPC to Intel-based vector instructions, are as fast as state-of-the-art implementations, yet much simpler. As an example, our straightforward implementation of string matching beats the Knuth-Morris-Pratt algorithm by 12%. This work was presented at the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP) 2017 [45].

This work was done in collaboration with Rubens Emilio and Fernando Pereira at UFMG, as part of the Inria PROSPIEL Associate Team.

6.1.7. Qubit allocation for quantum circuit compilers

Participant: Sylvain Collange.

Quantum computing hardware is becoming a reality. For instance, IBM Research makes a quantum processor available in the cloud to the general public. The possibility of programming an actual quantum device has elicited much enthusiasm. Yet, quantum programming still lacks the compiler support that modern programming languages enjoy today. To use universal quantum computers like IBM's, programmers must design low-level circuits. In particular, they must map logical qubits into physical qubits that need to obey connectivity constraints. This task resembles the early days of programming, in which software was built in machine languages. We have formally introduced the qubit allocation problem and provided an exact solution to it. This optimal algorithm deals with the simple quantum machinery available today; however, it cannot scale up to the more complex architectures scheduled to appear. Thus, we also provide a heuristic solution to qubit allocation, which is faster than the current solutions already implemented to deal with this problem.

This paper is accepted for publication at the Code Generation and Optimization (CGO) conference [49].

This work was done in collaboration with Vinícius Fernandes dos Santos, Fernando Pereira and Marcos Yukio Siraichi at UFMG, Brazil.

6.2. Processor Architecture

Participants: Pierre Michaud, Sylvain Collange, Erven Rohou, André Seznec, Biswabandan Panda, Fernando Endo, Kleovoulos Kalaitzidis, Daniel Rodrigues Carvalho, Anita Tino.

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

6.2.1. Microarchitecture

6.2.1.1. Bayesian TAGE predictors

Participant: Pierre Michaud.

The TAGE conditional branch predictor, introduced by André Seznec and Pierre Michaud in 2006, is the most storage-efficient branch predictor known today [19]. André Seznec has won the last four branch prediction championships, each time with a TAGE-based predictor. However, since 2006, the improvements in prediction accuracy have been relatively modest and were mostly obtained at the cost of increased hardware complexity. In particular, André Seznec added a Statistical Corrector to TAGE to correct some of its deficiencies [21]. This may be an indication that our understanding of TAGE is not complete and that further accuracy gains are waiting to be discovered. The problem tackled by the statistical corrector is that of cold counters: a TAGE-like predictor constantly allocate new entries, erasing the branch history information stored in the up-down counters of the overwritten entries. TAGE mitigates this problem by using the confidence level of the up-down counter and a meta-predictor. However, fundamentally, the information on the degree of coldness of the up-down counter is not available in TAGE. Therefore we propose to replace the up-down counter with a dual-counter counting separately taken and not-taken occurrences. Replacing the up-down counter with a dual-counter requires to redefine prediction confidence estimation. We found that a Bayesian formula, namely Laplace's rule of succession, provides effective confidence estimation. We also discovered a method, based on the dual-counter, for reducing the number of allocations. By combining these new findings, we devised a new TAGE-like predictor called BATAGE, more accurate than TAGE, making external statistical correction superfluous. As of December 2017, this work is in the process of being submitted to a journal.

6.2.1.2. Interactions Between Value Prediction and Compiler Optimizations
Participants: André Seznec, Fernando Endo.

Increasing instruction-level parallelism is regaining attractiveness within the microprocessor industry. The EOLE microarchitecture [13] and D-VTAGE value predictor [14] were recently introduced to solve practical issues of value prediction (VP). In particular, they remove the most significant difficulties that forbade an effective VP hardware. In [28], we present a detailed evaluation of the potential of VP in the context of EOLE/D-VTAGE and different compiler options. Our study shows that if no single general rule always applies – more optimization might sometimes leads to more performance – unoptimized codes often gets a large benefit from the prediction of redundant loads.

6.2.1.3. Prefetch Management on Multicore Systems
Participants: André Seznec, Biswabandan Panda.

In multi-core systems, an application's prefetcher can interfere with the memory requests of other applications using the shared resources, such as last level cache and memory bandwidth. Towards this end, we propose a solution to manage prefetching in multi-core systems [32]. In particular, we make two fundamental observations: First, a strong positive correlation exists between the accuracy of a prefetcher and the amount of prefetch requests it generates relative to an application's total (demand and prefetch) requests. Second, a strong positive correlation exists between the ratio of total prefetch to demand requests and the ratio of average last level cache miss service times of demand to prefetch requests. In [32], we propose Band-pass prefetching a simple and low-overhead mechanism to effectively manage prefetchers in multi-core systems that builds on those two observations. Our solution consists of local and global prefetcher aggressiveness control components, which altogether, control the flow of prefetch requests between a range of prefetch to demand requests ratios.

6.2.1.4. Managing Shared Last Level Caches in Large Multicores
Participant: André Seznec.

Multi-core processors employ shared Last Level Caches (LLC). This trend continuewith large multi-core processors (16 cores and beyond) as well. At the same time, the associativity of LLC tends to remain in the order of sixteen. Consequently, with large multicore processors, the number of applications or threads 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 applications. In [33], 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 applications. ADAPT builds on the use of the Footprint-number metric. 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. We further find that deprioritizing certain applications during cache replacement is beneficial to the overall performance.

6.2.1.5. Augmenting superscalar architecture for efficient many-thread parallel execution
Participants: Sylvain Collange, André Seznec.

Threads of Single-Program Multiple-Data (SPMD) applications often exhibit very similar control flows, i.e. they execute the same instructions on different data. We propose the Dynamic Inter-Thread Vectorization Architecture (DITVA) to leverage this implicit data-level parallelism in SPMD applications by assembling 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, multiple scalar threads running in lockstep share a single instruction stream and their respective instruction instances are aggregated into SIMD instructions. To balance thread-and data-level parallelism, threads are statically grouped into fixed-size independently scheduled warps. DITVA leverages existing SIMD units and maintains binary compatibility with existing CPU

architectures. Our evaluation on the SPMD applications from the PARSEC and Rodinia OpenMP benchmarks shows that a 4-warp × 4-lane 4-issue DITVA architecture with a realistic bank-interleaved cache achieves 1.55× higher performance than a 4-thread 4-issue SMT architecture with AVX instructions while fetching and issuing 51 % fewer instructions, achieving an overall 24 % energy reduction.

This work has been accepted for publication in the Journal of Parallel and Distributed Computing [30]. It was done in collaboaration with Sajith Kalathingal and Bharath Swamy from Intel Bangalore (India).

6.2.1.6. Generalizing the SIMT execution model to general-purpose instruction sets

Participant: Sylvain Collange.

The *Single Instruction, Multiple Threads* (SIMT) execution model as implemented in NVIDIA Graphics Processing Units (GPUs) associates a multi-thread programming model with an SIMD execution model [57]. It combines the simplicity of scalar code from the programmer's and compiler's perspective with the efficiency of SIMD execution units at the hardware level. However, current SIMT architectures demand specific instruction sets. In particular, they need specific branch instructions to manage thread divergence and convergence. Thus, SIMT GPUs have remained incompatible with traditional general-purpose CPU instruction sets.

We designed Simty, an SIMT processor proof of concept that lifts the instruction set incompatibility between CPUs and GPUs. Simty is a massively multi-threaded processor core that dynamically assembles SIMD instructions from scalar multi-thread code. It runs the RISC-V (RV32-I) instruction set. Unlike existing SIMD or SIMT processors like GPUs, Simty takes binaries compiled for general-purpose processors without any instruction set extension or compiler changes. Simty is described in synthesizable RTL. A FPGA prototype validates its scaling up to 2048 threads per core with 32-wide SIMD units.

The Simty architecture was presented at the First Workshop on Computer Architecture Research with RISC-V (CARRV 2017) [40].

Both conventional and generalized SIMT architectures like Simty use hardware or software mechanisms to keep track of control-flow divergence and convergence among threads. A new class of such algorithms is gaining popularity in the literature in the last few years. We presented a new classification of these techniques based on their common characteristic, namely traversals of the control-flow graph based on lists of paths. We compared the implementation cost on an FPGA of path lists and per-thread program counters within the Simty processor. The sorted list enables significantly better scaling starting from 8 threads per warp.

This work was presented in French in Conférence d'informatique en Parallélisme, Architecture et Système (ComPAS) [51] and is available in English as a technical report [52].

6.2.1.7. Toward out-of-order SIMT microarchitecture

Participants: Sylvain Collange, Anita Tino.

Prior work highlights the continued importance of maintaining adequate sequential performance within throughput-oriented cores [60]. Out-of-order superscalar architectures as used in high-performance CPU cores can meet such demand for single-thread performance. However, GPU architectures based on SIMT have been limited so far to in-order execution because of a major scientific obstacle: the partial dependencies between instructions that SIMT execution induces thwart register renaming. This ongoing project is seeking to generalize out-of-order execution to SIMT architectures. In particular, we revisit register renaming techniques originally proposed for predicate conversion to support partial register updates efficiently. Out-of-order dynamic vectorization holds the promise to close the CPU-GPU design space by enabling low-latency, high-throughput design points.

6.3. WCET estimation and optimization

Participants: Isabelle Puaut, Damien Hardy, Viet Anh Nguyen, Benjamin Rouxel, Sébastien Martinez, Erven Rohou, Imen Fassi, Loïc Besnard, Stefanos Skalistis.

6.3.1. WCET estimation for many core processors

Participants: Viet Anh Nguyen, Damien Hardy, Sébastien Martinez, Isabelle Puaut, Benjamin Rouxel.

6.3.1.1. Optimization of WCETs by considering the effects of local caches

The overall goal of this research is to define 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.

We have addressed this issue by developing both optimal and heuristic techniques for solving the scheduling problem, whose objective is to minimize the WCET of a parallel application. Our proposed static partitioned non-preemptive mapping strategies address the effect of local caches to tighten the estimated WCET of the parallel application. Experimental results obtained on real and synthetic parallel applications show that co-locating tasks that reuse code and data improves the WCET by 11 % on average for the optimal method and by 9 % on average for the heuristic method [35].

This research is part of the PIA Capacités project.

6.3.1.2. Accounting for shared resource contentions to minimize WCETs

Accurate WCET analysis for multi-cores is known to be challenging, because of concurrent accesses to shared resources, such as communication through busses or Networks on Chips (NoC). Since it is impossible in general to guarantee the absence of resource conflicts during execution, current WCET techniques either produce pessimistic WCET estimates or constrain the execution to enforce the absence of conflicts, at the price of a significant hardware under-utilization. In addition, the large majority of existing works consider that the platform workload consists of independent tasks. As parallel programming is the most promising solution to improve performance, we envision that within only a few years from now, real-time workloads will evolve toward parallel programs. The WCET behavior of such programs is challenging to analyze because they consist of *dependent* tasks interacting through complex synchronization/communication mechanisms.

In a first work (thesis of Benjamin Rouxel), we proposed techniques that account for interferences to access shared ressources, in order to minimize the WCET of parallel applications. An optimal and a heuristic method are proposed to map and schedule tasks on multi-cores. These methods take the structure of applications (synchronizations/communications) into consideration to tightly identify shared resource interferences and consequently tighten WCET estimates. Our heuristic improves by 19% the overall WCET compared to a worst-case contention baseline [47], [31].

In a second study [44], we have studied the gain that could be obtained on an initially produced time-triggered non-preemptive schedule, by the introduction of slack time, in order to avoid interference between tasks. The introduction of slack time is performed using an optimal technique using Integer Linear Programming (ILP), to evaluate how much at best can be gained. Experimental results using synthetic task graphs and a Kalray-like architecture with round-robin bus arbitration show that avoiding contention reduces WCETs, albeit by a small percentage. The highest reductions are observed on applications with the highest memory demand, and when the application is scheduled on the highest number of cores.

This work is performed in cooperation with Steven Derrien from the CAIRN research group and is part of the ARGO H2020 project.

6.3.1.3. WCET-Aware Parallelization of Model-Based Applications for Multi-Cores

Parallel architectures are nowadays no longer confined to the domain of high performance computing, they are also increasingly used in embedded time-critical systems.

The ongoing ARGO H2020 project provides a programming paradigm and associated tool flow to exploit the full potential of architectures in terms of development productivity, time-to-market, exploitation of the platform computing power and guaranteed real-time performance. In [41] we give an overview of the objectives of ARGO and explore the challenges introduced by our approach.

6.3.2. WCET estimation tool and benchmarks

Participants: Damien Hardy, Isabelle Puaut, Benjamin Rouxel, Loïc Besnard.

Estimation of worst-case execution times (WCETs) is required to validate the temporal behavior of hard real time systems. Heptane is an open-source software program that estimates upper bounds of execution times on MIPS and ARM v7 architectures, offered to the WCET estimation community to experiment new WCET estimation techniques. The software architecture of Heptane was designed to be as modular and extensible as possible to facilitate the integration of new approaches. In [42], we present the current status of Heptane, give information on the analyses it implements, as well as how to use it and extend it.

We all had quite a time to find non-proprietary architecture-independent exploitable parallel benchmarks for Worst-Case Execution Time (WCET) estimation and real-time scheduling. However, there is no consensus on a parallel benchmark suite, when compared to the single-core era and the Mälardalen benchmark suite. In [48] we bridge part of this gap, by presenting a collection of benchmarks with the following good properties: (i) easily analyzable by static WCET estimation tools (written in structured C language, in particular neither goto nor dynamic memory allocation, containing flow information such as loop bounds); (ii) independent from any particular run-time system (MPI, OpenMP) or real-time operating system. Each benchmark is composed of the C source code of its tasks, and an XML description describing the structure of the application (tasks and amount of data exchanged between them when applicable). Each benchmark can be integrated in a full end-to-end empirical method validation protocol on multi-core architecture. This proposed collection of benchmarks is derived from the well known StreamIT benchmark suite and will be integrated in the TACleBench suite in a near future.

6.4. Security

Participants: Erven Rohou, Damien Hardy, Nicolas Kiss.

Physical attacks represent a very important threat in the context of embedded systems: these attacks try to recover cryptographic keys by exploiting the physical behavior of the device. They can either be passive (e.g. by monitoring the power consumption of the device) or active (e.g. by injecting errors to reveal or deduce sensitive data).

One family of countermeasures to protect against those passive attacks (also known as *side-channel* attacks) is called masking. The principle is to “hide” data with masks so that internal values used in computations can not be predicted with the behavior observed. We modified the LLVM compiler (version 3.8) to automatically insert masking coutermeasures into the code at compile-time. Our modification works at intermediate level (IR level), this way we can perform low-level transformations (e.g. memory allocation, instructions replacement) while covering most of the architectures used in the embedded world.

The main innovation of this work is the generic approach used for the transformation and thus, the ability to easily change the masking scheme without modifying the compiler internal code. We introduced a way to describe in high-level language (C/C++) the masking operations independently in what we call “primitives”. With this technique, we implemented “Boolean Masking” and we tested the efficiency on an embedded implementation of AES. After measuring the electromagnetic emissions of 20,000 executions, we performed a Correlation Power Analysis (CPA) and results have shown that the countermeasure is correctly applied. Hence, it is not possible anymore to recover the cryptographic key with this type of attack.

This work is done in the context of the SECODE CHIST-ERA project.

SUMO Project-Team

7. New Results

7.1. Analysis and Verification of Quantitative Systems

7.1.1. Diagnosability

Participants : Hugo Bazille, Éric Fabre, Blaise Genest, Loïc Hélouët, Hervé Marchand, Engel Lefaucheux

7.1.1.1. Diagnosability of repairable faults.

Diagnosability (i.e., the existence of a diagnoser detecting faults in partially-observable systems) can be decided in polynomial time, relying on the so-called twin-machine construction. We have examined the case of repairable faults, and a notion of diagnosability that requires the detection of the fault before it is repaired. We have extended a contribution of 2016 to show that diagnosability of faults and of their repair could help counting the number of occurred faults. It was proved [51] that diagnosability with repair is a *PSPACE*-complete problem. We have completed this result, showing that the close notion of P-diagnosability (diagnosability of a fault even after it is repaired) is also *PSPACE*-complete [20].

7.1.1.2. Diagnosability degree of stochastic systems.

For stochastic systems, several diagnosability properties have been defined. The simplest one, also called A-diagnosability, characterizes the fact that after each fault, detection will almost surely occur. We have considered quantitative versions of the problem, to determine how much a system is diagnosable (when it is not diagnosable for sure). This amounts to characterizing the probability that a faulty run will lead to detection. We have proposed several notions of dignosability degree. Their derivation is generally *NP*-hard, but we have identified situations where complexity becomes polynomial. Besides, we have developed techniques to compute the different moments of the detection delay (mean, variance and upper moments). This allows one to compare systems with similar detection degrees, but that can react faster to faults. In some cases, one may be able to tune a system and trade diagnosability degree againts a faster detection. This approach also yields the distribution of fault location (in time) once detection takes place. Given the first moments of the detection delay, one is also able to compute (sometimes tight) bounds on the response time, for example to lower bound the probability that detection takes place at most T seconds/events after the fault [31].

7.1.1.3. The cost of diagnosis.

We addressed diagnosability and its cost for safe Petri nets. In [37] we have defined an energy-like cost model for Petri nets: transitions can consume or restore energy of the system. We then have defined a partial-order representation for state estimation, and extend the cost model and the capacities of diagnosers. Diagnosers are allowed to use additional energy to refine their estimations. Diagnosability is then seen as an energy game: checking whether disambiguation mechanisms are sufficient to allow diagnosability is in *2EXPTIME*, and one can also decide in *2EXPTIME* whether diagnosability under budget constraint holds.

7.1.2. Analysis of timed systems

Participants : Nicolas Markey, Loïc Hélouët

7.1.2.1. Determinizing timed automata.

In [35], we introduce a new formalism called *automata over a timed domain*, which generalizes timed automata; this formalism provides an adequate framework for determinization. In our formalism, determinization w.r.t. timed language is always possible at the cost of changing the timed domain. We give a condition for determinizability of automata over a timed domain *without changing the timed domain*, which allows us to recover several known determinizable classes of timed systems, such as strongly-non-zeno timed automata, integer-reset timed automata, perturbed timed automata, etc. Moreover, in the case of timed automata, this condition encompasses most determinizability conditions from the literature. Our aim now is to extend this work towards more efficient algorithms for monitoring timed systems.

7.1.2.2. Concurrent Timed Systems.

Time Petri nets (TPNs) are a classical extension of Petri nets with timing constraints attached to transitions, for which most verification problems are undecidable. We consider TPNs under a strong semantics with multiple enablings of transitions. This year, we have extened a work started in 2016, focusing on a structural subclass of unbounded TPNs, where the underlying untimed net is free choice, and showed that it enjoys nice properties in the timed setting under a multi-server semantics [46], [25]. In particular, we have showed that the questions of firability (whether a chosen transition can fire), and termination (whether the net has a non-terminating run) are decidable for this class. Next, we have considered the problem of robustness under guard enlargement and guard shrinking, i.e., whether a given property is preserved even if the system is implemented on an architecture with imprecise time measurement. For unbounded free choice TPNs with a multi-server semantics, we have show decidability of robustness of firability and of termination under both guard enlargement and shrinking.

7.2. Control of Quantitative Systems

7.2.1. Expressing and verifying properties of multi-agent systems

Participants : Ocan Sankur, Nicolas Markey

7.2.1.1. Admissible strategies in controller synthesis.

In game theory, a strategy is dominated by another one if the latter systematically yields a payoff as good as the former, while also yielding a better payoff in some cases. A strategy is admissible if it is not dominated. This notion is well-studied in game theory and is useful to describe the set of strategies that are "reasonable" (i.e., whose choice can be justified; here, no players would play a dominated strategy, since better strategies exist). Recent works studied this notion in graph games with omega-regular objectives and investigated its applications in controller synthesis. For multi-agent controller synthesis, admissibility can be used as a hypothesis on the behaviors of each agent, thus enabling a compositional reasoning framework for controller synthesis.

We continue the study of admissibility in controller synthesis with three developments detailed as follows:

- In [29], we study the characterization and computation of admissible strategies in multiplayer concurrent games. We study both deterministic strategies and randomized ones with almost-sure winning criteria. We prove that admissible strategies always exist in concurrent games, and we characterise them precisely. Then, when the objectives of the players are omega-regular, we show how to perform assume-admissible synthesis, i.e., how to compute admissible strategies that win (almost surely) under the hypothesis that the other players play admissible strategies only.
- In [30], we study timed games, which are multiplayer games played on arena defined by timed automata, which are a particular case of concurrent games. First, we show that admissible strategies may not exist in timed games with a continuous semantics of time, even for safety objectives. Second, we show that the discrete time semantics of timed games is better behaved w.r.t. admissibility: the existence of admissible strategies is guaranteed in that semantics. Third, we provide symbolic algorithms to solve the model-checking problem under admissibility and the assume-admissible synthesis problem for real-time non-zero sum n-player games for safety objectives.
- In [26], we study admissible strategies in games with imperfect information. We show that in stark contrast with the perfect information variant, admissible strategies are only guaranteed to exist when players have objectives that are closed sets. As a consequence, we also study decision problems related to the existence of admissible strategies for regular games as well as finite duration games.

7.2.1.2. Strategy dependences in Strategy Logic.

Strategy Logic (SL) is a very expressive logic for specifying and verifying properties of multi-agent systems: in SL, one can quantify over strategies, assign them to agents, and express properties of the resulting plays (using linear-time temporal logic). This defines a very expressive framework, encompassing e.g. (pure) Nash equilibria, or admissibility. Such a powerful framework has two drawbacks: first, SL model checking has non-elementary complexity; second, the exact semantics of SL is rather intricate, and may not correspond to what is expected.

In [49], we focus on *strategy dependences* in SL, by tracking how existentially-quantified strategies in a formula may (or may not) depend on other strategies selected in the formula. We study different kinds of dependences, refining a previous approach [52], and prove that they give rise to different satisfaction relations. In the setting where strategies may only depend on what they have observed, we identify a large fragment of SL for which we prove model checking can be performed in *2EXPTIME*.

7.2.2. Active diagnosis

Participants : Nathalie Bertrand, Blaise Genest, Engel Lefaucheux

7.2.2.1. Diagnosis and control of the degradation of probabilistic systems.

Active diagnosis is performed by a controler so that a system becomes diagnosable. In order to avoid the controler to degrade the funtionning of the system too much, one often provides it with an additional objective specifying the desired quality of service.

In the context of probabilistic systems, a possible specification consists in requiring a positive probability of infinite correct runs, referred to as the safe active diagnosis. In [42], we introduced two alternative specifications. First (γ, v)-correction of a system associates with an execution a correction value which depends on a discount factor γ, and the controler must ensure an expected correction value greater than a threshold v. Second, α-persistence requires that asymptotically, at each time unit, a proportion at least α of runs that were correct so far remain correct.

Our contributions are twofold. On the one hand, from a semantical viewpoint, we make explicit the equivalences and (non-)implications between the various notions, for finite-state systems as well as infinite-state ones. On the other hand, algorithmically, we establish the decidability frontier of the corresponding decision problems, and for decidable problems characterize their precise complexity, together with algorithms to design controlers.

7.2.2.2. Probabilistic Disclosure: Maximisation vs. Minimisation.

We consider opacity questions where an observation function provides to an external attacker a view of the states along executions and secret executions are those visiting some secret state from a fixed subset. Disclosure occurs when the observer can deduce from a finite observation that the execution is secret. In a probabilistic and non deterministic setting, where an internal agent can choose between actions, there are two points of view, depending on the status of this agent: the successive choices can either help the attacker trying to disclose the secret, if the system has been corrupted, or they can prevent disclosure as much as possible if these choices are part of the system design. In the former situation, corresponding to a worst case, the disclosure value is the supremum over the strategies of the probability to disclose the secret (maximisation), whereas in the latter case, the disclosure is the infimum (minimisation). We address quantitative problems (relation between the optimal value and a threshold) and qualitative ones (when the threshold is zero or one) related to both forms of disclosure for a fixed or finite horizon. For all problems, we characterise their decidability status and their complexity. Surprisingly, while in maximisation problems optimal strategies may be chosen among deterministic ones, it is not the case for minimisation problems, but more minimisation problems than maximisation ones are decidable. These results appeared in [36].

7.2.3. Control and enforcement for quantitative systems

Participants : Nathalie Bertrand, Blaise Genest, Thierry Jéron, Hervé Marchand, Nicolas Markey

7.2.3.1. Qualitative determinacy and Decidability of Stochastic Games with Signals.

In [17], we consider two-person zero-sum stochastic games with signals, a standard model of stochastic games with imperfect information. The only source of information for the players consists of the signals they receive; they cannot directly observe the state of the game, nor the actions played by their opponent, nor their own actions.

We are interested in the existence of almost-surely winning or positively winning strategies, under reachability, safety, Büchi, or co-Büchi winning objectives, and the computation of these strategies when the game has finitely many states and actions. We prove two qualitative determinacy results. First, in a reachability game, either player 1 can achieve almost surely the reachability objective, or player 2 can achieve surely the dual safety objective, or both players have positively winning strategies. Second, in a Büchi game, if player 1 cannot achieve almost surely the Büchi objective, then player 2 can ensure positively the dual co-Büchi objective. We prove that players only need strategies with finite memory. The number of memory states needed to win with finite-memory strategies ranges from one (corresponding to memoryless strategies) to doubly exponential, with matching upper and lower bounds. Together with the qualitative determinacy results, we also provide fix-point algorithms for deciding which player has an almost-surely winning or a positively winning strategy and for computing an associated finite-memory strategy. Complexity ranges from *EXPTIME* to *2EXPTIME*, with matching lower bounds. Our fix-point algorithms also enjoy a better complexity in the cases where one of the players is better informed than their opponent.

Our results hold even when players do not necessarily observe their own actions. The adequate class of strategies, in this case, is mixed or general strategies (they are equivalent). Behavioral strategies are too restrictive to guarantee determinacy: it may happen that one of the players has a winning general strategy but none of them has a winning behavioral strategy. On the other hand, if a player can observe their actions, then general, mixed, and behavioral strategies are equivalent. Finite-memory strategies are sufficient for determinacy to hold, provided that randomized memory updates are allowed.

7.2.3.2. Average-energy games.

In [34], we consider average-energy games, where the goal is to minimize the long-run average of the accumulated weight (seen as an *energy level*) in a two-player game on a finite-state weighted automaton. Decidability of average-energy games with a lower-bound constraint on the energy level (but no upper bound) is an open problem; in particular, there is no known upper bound on the memory that is required for winning strategies.

By reducing average-energy games with lower-bounded energy to infinite-state mean-payoff games and analyzing the frequency of low-energy configurations, we show an almost tight doubly-exponential upper bound on the necessary memory, and that the winner of average-energy games with lower-bounded energy can be determined in doubly-exponential time. We also prove *EXPSPACE*-hardness of this problem.

Finally, we consider multi-dimensional extensions of all types of average-energy games: without bounds, with only a lower bound, and with both a lower and an upper bound on the energy. We show that the fully-bounded version is the only case to remain decidable in multiple dimensions.

7.2.3.3. Runtime enforcement.

The journal paper [23] details our work about predictive runtime enforcement, done in collaboration with University Aalto (Finland) and Inria CORSE/LIG Grenoble.

Runtime enforcement (RE) is a technique to ensure that the (untrustworthy) output of a black-box system satisfies some desired properties. In RE, the output of the running system, modeled as a sequence of events, is fed into an enforcer. The enforcer ensures that the sequence complies with a certain property, by delaying or modifying events if necessary. This paper deals with 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 standard (non-predictive) runtime enforcement. All our results related to predictive RE of untimed properties are also formalized and proved in the Isabelle theorem prover. We also discuss how our predictive runtime enforcement framework can be extended to enforce timed properties.

The journal paper [24], done in collaboration with LaBRI Bordeaux and Inria Corse/LIG Grenoble, deals with runtime enforcement of untimed and timed properties with uncontrollable events. Runtime enforcement consists in defining and using mechanisms that modify the executions of a running system to ensure their

correctness with respect to a desired property. We introduce a framework that takes as input any regular (timed) property described by a deterministic automaton 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 should satisfy important properties, namely soundness, compliance, and optimality—meaning that enforcement mechanisms should output as soon as possible correct executions that are as close as possible to the input execution. We define the conditions for a property to be enforceable with uncontrollable events. Moreover, we synthesise sound, compliant, and optimal descriptions of runtime enforcement mechanisms at two levels of abstraction to facilitate their design and implementation.

7.2.3.4. Control of logico-numerical systems.

In paper [32], we have targeted the problem of the safe control of reconfigurations in component-based software systems, where strategies of adaptation to variations in both their environment and internal resource demands need to be enforced. In this context, the computing system involves software components that are subject to control decisions. We have approached this problem under the angle of discrete-event systems (DES), involving properties on events observed during the execution (e.g., requests of computing tasks, work overload), and a state space representing different configurations such as activity or assemblies of components. We have considered in particular the potential of applying novel logico-numerical control techniques to extend the expressivity of control models and objectives, thereby extending the application of DES in component-based software systems. We elaborate methodological guidelines for the application of logico-numerical control based on a case-study, and validate the result experimentally.

7.2.4. Smart regulation for urban trains

Participants : Éric Fabre, Loïc Hélouët, Hervé Marchand, Karim Kecir

The regulation of subway lines consists in accomodating small random perturbations in transit times as well as more impacting incidents, by playing on continuous commands (transit times and dwell times) and by making more complex decisions (insertions or extractions of trains, changes of missions, overpassing, shorter returns, etc.) The objectives are multiple: ensuring the regularity and punctuality of trains, adapting to transportation demand, minimizing energy consumption, etc. We have developed an event-based control strategy that aims at equalizing headways on a line. This distributed control strategy is remarquably robust to perturbations and reactive enough to accomodate train insertions/extractions. We have integrated this control startegy to our SIMSTORS software. We have also developed another approach based on event graphs in order to optimally interleave trains at a junction. We started investigating new predictive control policies based of optimisation of criteria in forecast schedules [43].

In [47], we have extended a work started in 2016, that considers realizability of schedules by metro systems. Schedules are defined as high-level views of desired executions of systems, and represented as partial orders decorated with timing constraints. Train networks are modeled as stochastic time Petri nets (STPN) with an elementary (1-bounded) semantics. We have proposed a notion of time processes to give a partial-order semantics to STPNs. We then have considered Boolean realizability: a schedule S is realizable by a net N if S embeds in a time process of N that satisfies all its constraints. However, with continuous time domains, the probability of a time process with exact dates is null. We thus consider probabilistic realizability up to α time units, that holds if the probability that N realizes S with constraints enlarged by α is strictly positive. Upon a sensible restriction guaranteeing time progress, Boolean and probabilistic realizability of a schedule can be checked on the finite set of symbolic prefixes extracted from a bounded unfolding of the net. We give a construction technique for these prefixes and show that they represent all time processes of a net occurring up to a given maximal date. We then show how to verify existence of an embedding and compute the probability of its realization. The technique has then been illustrated by a concrete example, namely deciding wheter a simple flip-flop shunting mechanism suffices to route trains in appropriate direction when delays can occur in trips or during stops at stations. We have also conducted a series of experiment [28] with the SIMSTORS tool to obtain statistics, and show feasibiliy of Key Performance Indicators (KPIs) evaluation with this formal model.

A second line of research relates to the development of new regulation strategies. New techniques were derived to equalize headways of trains along a line, and thus improve regularity and resilience to perturbations. A distributed control strategy was developed, easily implementable in existing rule engines. Simulations have proved the efficiency of this technique on orbital lines. We have also developed a global regulation approach based on timed event graphs. In this setting, control is event-based: a command is issued each time a train crosses a control point, but it takes into account information along the whole line and for a finite time horizon. This amounts to adapting the whole time-table for any new event in the system. This approach has been proved to perform well at junctions (on computer simulations), where randomly spaced trains arriving from two branches must be correctly interleaved at the junction of the two lines, while at the same time train intervals must be equalized in all branches. We are now working on the combinatorial aspects of the question, in order to reduce energy consumption (by synchronizing arrivals and departures of trains), and in order to allow for insertions/extractions and reorderings of trains.

Several patents are in preparation for this activity.

7.3. Management of Large Distributed Systems

7.3.1. Analysis and synthesis of distributed systems

Participants : Éric Badouel, Thierry Jéron, Hervé Marchand, The Anh Pham

7.3.1.1. Control of Distributed Systems.

In [40], we have extended our examination of decentralized discrete-event-system architectures that use exclusive or (XOR) as the fusion rule to reach control decisions. A characterization of XOR inference-observable languages has been provided. Additionally, XOR observability is defined for languages that are not inference-observable but are distributed-observable.

7.3.1.2. Verification of distributed applications

In the context of IPL HAC-SPECIS, in collaboation with Martin Quinson (Myriads Inria project team) we are interested in the verification of real distributed applications.

In the conference paper [38] we explain the current status of the tool SimGridMC used for the verification of MPI applications. SimGridMC (also dubbed Mc SimGrid) is a stateful Model Checker for MPI applications. It is integrated to SimGrid, a framework mostly dedicated to predicting the performance of distributed applications. We describe the architecture of McSimGrid, and show how it copes with the state space explosion problem using Dynamic Partial Order Reduction and State Equality algorithms. As case studies we show how SimGrid can enforce safety and liveness properties for MPI applications, as well as global invariants over communication patterns.

7.3.2. Analysis of parameterized systems

Participants : Nathalie Bertrand, Éric Fabre, Blaise Genest, Matthieu Pichené, Ocan Sankur

7.3.2.1. Parameterized Verification of a time-synchronization protocol.

In [41], we consider distributed timed systems that implement leader-election protocols, which are at the heart of clock-synchronization protocols. We develop abstraction techniques for parameterized model checking of such protocols under arbitrary network topologies, where nodes have independently-evolving clocks. We apply our technique for model checking the root election part of the flooding time-synchronisation protocol (FTSP), and obtain improved results compared to previous work. We model-check the protocol for all topologies in which the distance to the node to be elected leader is bounded by a given parameter.

7.3.2.2. Controlling population models.

In [33], we introduce a new setting where a population of agents, each modelled by a finite-state system, are controlled uniformly: the controller applies the same action to every agent. The framework is largely inspired by the control of a biological system, namely a population of yeasts, where the controller may only change the environment common to all cells. We study a synchronisation problem for such populations: no matter how individual agents react to the actions of the controller, the controller aims at driving all agents synchronously to a target state. The agents are naturally represented by a non-deterministic finite state automaton (NFA), the same for every agent, and the whole system is encoded as a 2-player game. The first player (Controller) chooses actions, and the second player (Agents) resolves non-determinism for each agent. The game with m agents is called the m-population game. This gives rise to a parameterized control problem (where control refers to 2-player games), namely the population control problem: can Controller control the m-population game for all $m \in \mathbb{N}$, whatever Agents does?

In this work, we prove that the population control problem is decidable, and it is an *EXPTIME*-complete problem. As far as we know, this is one of the first results on parameterized control. Our algorithm, not based on cut-off techniques, produces winning strategies which are symbolic, that is, they do not need to count precisely how the population is spread between states. We also show that if there is no winning strategy, then there is a population size M such that Controller wins the m-population game if, and only if, $m \leq M$. Surprisingly, M can be doubly-exponential in the number of states of the NFA, with tight upper and lower bounds.

7.3.2.3. Handling large biological systems.

This year, we propose to use approximated probabilistic distribution to handle large homogeneous populations of cells [39]. Beyond classical approximations, we propose to use the Chow-Liu tree representation, based on *non-disjoint* clusters of two variables. Our experiments show that our proposed approximation scheme is more accurate than existing ones to model probability distributions deriving from biopathways, while requiring a minimal complexity overhead.

To handle *dynamics* of a population of cells governed by biopathways, we develop *coarse-grained* abstractions of the biological pathways [21], and more precisely *Dynamic Bayesian Networks* (DBNs). We show that simulating a DBN is much faster than simulating the fine-grained model it abstracts, for comparable prediction performances.

We also explore the approximate inference problem of DBNs, that is, *computing* the probability distributions at every time point given the initial distribution at time 0. We evaluate several classical approximate inference algorithms for DBNs, and compare with a new method we propose, which consists in using the Chow-Liu tree approximation to represent distributions at each time step. It is very accurate, yet efficient according to experiments we report. We finally provide an error analysis of this approximate inference algorithm [39].

7.4. Data-Driven Systems

7.4.1. Incremental process discovery using Petri-net synthesis.

Participants : Éric Badouel

In [16], we present an incremental process discovery using Petri-net synthesis. Process discovery aims at constructing a model from a set of observations given by execution traces (a log). Petri nets are a preferred target model in that they produce a compact description of the system by exhibiting its concurrency. This article presents a process-discovery algorithm using Petri-net synthesis, based on the notion of region introduced by A. Ehrenfeucht and G. Rozenberg, and using techniques from linear algebra. The algorithm proceeds in three successive phases which make it possible to find a compromise between the ability to infer behaviours of the system from the set of observations while ensuring a parsimonious model, in terms of fitness, precision and simplicity. All used algorithms are incremental which means that one can modify the produced model when new observations are reported without reconstructing the model from scratch.

7.4.2. An artifact model with imprecision and uncertainty

Participants : Éric Badouel, Loïc Hélouët

In the context of the HeadWork ANR project, we started investigating how complex workflows can be defined to handle uncertainty, and use joint knowledge of pools of user to build correct information. The solution proposed so far is a variant of business artifact managing fuzzy datasets. As there are several ways to reach an acceptable final and sufficiently precise dataset, we started investigating equivalence of complex workflows with partial information to allow refinement, enhance performance of data collection, with mastered precision loss.

TAMIS Team

7. New Results

7.1. Results for Axis 1: Vulnerability analysis

7.1.1. Statistical Model Checking of LLVM Code

Participants: Axel Legay, Louis-Marie Traonouez.

We have extended PLASMA Lab statistical model-checker with a new plugin that allows to simulate LLVM bitcode. The plugin is based on an external simulator LODIN. This simulator implements a probabilistic semantics for a LLVM program. At its core the semantics consist of the LLVM program given as a labelled transition system. The labels are function calls to an environment that implements functions outside the LLVM core language. The environment is also responsible for assigning probabilities to individual transitions

By interfacing the LODIN simulator with PLASMA Lab we can apply all the statistical model-checking algorithms provided by PLASMA Lab, including rare events verification algorithms like importance splitting. We have applied LODIN and PLASMA Lab to several case studies, including the analysis of some security vulnerability, like the PTrace privilege escalation attack that could be performed on earlier versions of the Linux Kernel. This work has been submitted to a conference this year [61], and is currently under review.

[61] We present our work in providing Statistical Model Checking for programs in LLVM bitcode. As part of this work we develop a semantics for programs that separates the program itself from its environment. The program interact with the environment through function calls. The environment is furthermore allowed to perform actions that alter the state of the C-program-useful for mimicking an interrupt system. On top of this semantics we build a probabilistic semantics and present an algorithm for simulating traces under that semantics. This paper also includes the development of the new tool component Lodin that provides a statistical model checking infrastructure for LLVM programs. The tool currently implement standard Monte Carlo algorithms and a simulator component to manually inspect the behaviour of programs. The simulator also proves useful in one of our other main contributions; namely producing the first tool capable of doing importance splitting on LLVM code. Importance splitting is implemented by integrating Lodin with the existing statistical model checking tool Plasma-Lab.

7.1.2. Verification of IKEv2 protocol

Participants: Axel Legay, Tristan Ninet, Louis-Marie Traonouez, Olivier Zendra.

The IKEv2 (Internet Key Exchange version 2) protocol is the authenticated key-exchange protocol used to set up secure communications in an IPsec (Internet Protocol security) architecture. It guarantees security properties like mutual-authentication and secrecy of the exchanged key. To obtain an IKEv2 implementation as secure as possible, we use model checking to verify the properties on the protocol specification, and smart fuzzing to test the implementation, and try to detect implementation flaws like buffer overflows or memory leaks.

Two weaknesses had previously been found in the specification, but were harmless. We showed that the first weakness does not actually exist. We demonstrated that the second weakness is not harmless, and we designed a Denial-of-Service attack that exploits it, the deviation attack. As a counter-measure, we propose a modification of IKEv2, and use model checking to prove that the modified version is secure.

This work is being prepared for responsive disclosure and publication.

7.1.3. High-Level Frameworks for Scheduling Systems

Participants: Mounir Chadli, Axel Legay, Louis-Marie Traonouez.

Formal model-based techniques are more and more used for the specification and verification of scheduling systems. These techniques allow to consider complex scheduling policies beyond the scope of classical analytical techniques. For instance, hierarchical scheduling systems (HSS) integrates a number of components into a single system running on one execution platform. Hierarchical scheduling systems have been gaining more attention by automotive and aircraft manufacturers because they are practical in minimizing the cost and energy of operating applications. Model-based techniques can also be used to solve new problems like energy optimization or runtime monitoring. However, one limitation of formal model-based approaches is that they require high technical knowledge about the formalims and tools used to design models and write properties.

In a previous work [62], we have presented a model-based framework for the verification of HSS. It is based on a stochastic extension of timed automata and statistical model checking with the tool UPPAAL. We have also developed a graphical high-level language to represent complex hierarchical scheduling systems. To bridge the gap between the formalisms, we exploit Cinco, a generator for domain specific modeling tools to generate an interface between this language and the one of UPPAAL. Cinco allows to specify the features of a graphical interface in a compact meta-model language. This is a flexible approach that could be extended to any formal model of scheduling problem.

We have extended the previous work in journal paper [55] published this year, where we provide another high-level framework for the verification of energy-aware scheduling systems. We also present two new analysis techniques. One that performs runtime monitoring in order to detect alarming change in the scheduling system, and one that performs energy optimization.

[55] Over the years, schedulability of Cyber-Physical Systems (CPS) has mainly been performed by analytical methods. These techniques are known to be effective but limited to a few classes of scheduling policies. In a series of recent work, we have shown that schedulability analysis of CPS could be performed with a model-based approach and extensions of verification tools such as UPPAAL. One of our main contributions has been to show that such models are flexible enough to embed various types of scheduling policies, which goes beyond those in the scope of analytical tools.

However, the specification of scheduling problems with model-based approaches requires a substantial modeling effort, and a deep understanding of the techniques employed in order to understand their results. In this paper we propose simplicity-driven high-level specification and verification frameworks for various scheduling problems. These frameworks consist of graphical and user-friendly languages for describing scheduling problems. The high-level specifications are then automatically translated to formal models, and results are transformed back into the comprehensible model view. To construct these frameworks we exploit a meta-modeling approach based on the tool generator Cinco.

Additionally we propose in this paper two new techniques for scheduling analysis. The first performs runtime monitoring using the CUSUM algorithm to detect alarming change in the system. The second performs optimization using efficient statistical techniques. We illustrate our frameworks and techniques on two case studies.

7.1.4. Side-channel Analysis of Cryptographic Substitution Boxes

Participants: Axel Legay, Annelie Heuser.

With the advent of the Internet of Things, we are surrounded with smart objects (aka things) that have the ability to communicate with each other and with centralized resources. The two most common and widely noticed artefacts are RFID and Wireless Sensor Networks which are used in supply-chain management, logistics, home automation, surveillance, traffic control, medical monitoring, and many more. Most of these applications have the need for cryptographic secure components which inspired research on cryptographic algorithms for constrained devices. Accordingly, lightweight cryptography has been an active research area over the last 10 years. A number of innovative ciphers have been proposed in order to optimize various performance criteria and have been subject to many comparisons. Lately, the resistance against side-channel attacks has been considered as an additional decision factor.

Side-channel attacks analyze physical leakage that is unintentionally emitted during cryptographic operations in a device (e.g., power consumption, electromagnetic emanation). This side-channel leakage is statistically dependent on intermediate processed values involving the secret key, which makes it possible to retrieve the secret from the measured data.

Side-channel analysis (SCA) for lightweight ciphers is of particular interest not only because of the apparent lack of research so far, but also because of the interesting properties of substitution boxes (S-boxes). Since the nonlinearity property for S-boxes usually used in lightweight ciphers (i.e., 4×4) can be maximally equal to 4, the difference between the input and the output of an S-box is much smaller than for instance for AES. Therefore, one could conclude that from that aspect, SCA for lightweight ciphers must be more difficult. However, the number of possible classes (e.g., Hamming weight (HW) or key classes) is significantly lower, which may indicate that SCA must be easier than for standard ciphers. Besides the difference in the number of classes and consequently probabilities of correct classification, there is also a huge time and space complexity advantage (for the attacker) when dealing with lightweight ciphers.

In [65], [64] we give a detailed study of lightweight ciphers in terms of side-channel resistance, in particular for software implementations. As a point of exploitation we concentrate on the non-linear operation (S-box) during the first round. Our comparison includes SPN ciphers with 4-bit S-boxes such as KLEIN, PRESENT, PRIDE, RECTANGLE, Mysterion as well as ciphers with 8-bit S-boxes: AES, Zorro, Robin. Furthermore, using simulated data for various signal-to-noise ratios (SNR) we present empirical results for Correlation Power Analysis (CPA) and discuss the difference between attacking 4-bit and 8-bit S-boxes.

An extension of this work is given in [10]. We investigate whether side-channel analysis is easier for lightweight ciphers than e.g. for AES. We cover both profiled and non-profiled techniques where we are interested in recovering secret (round)keys or intermediate states. In the case of non-profiled attacks, we evaluate a number of S-boxes appearing in lightweight ciphers using the confusion coefficient and empirical simulations.

First, we investigate in the scenario where the attacker targets the first round and thus exploits the S-box computation. We observe that the 8-bit S-boxes from AES, Zorro, and Robin perform similarly, whereas for 4-bit S-boxes we have a clear ranking, with the S-box of Piccolo being the weakest to attack and the S-box of KLEIN and Midori (1) the hardest. Interestingly, when considering the last round and thus the inverse S-box operation the ranking changes such that Mysterion is the weakest and PRESENT/LED is the most side-channel resistant cipher from the ones investigated. Moreover, we could observe that attacking the last round is equal or less efficient for all considered ciphers. Finally, we use the information gained from both rounds together, where this approach is of interest when the cipher does not use round keys from a key scheduling algorithm but rather uses the same (or a straightforward computable) key in each round. LED fulfils this requirement. For a reasonable low SNR, to reach a success rate of 0.9 an attack on both rounds only requires 100 traces, whereas an attack using the first round requires 200 traces and on the last 400 traces. This example highlights the important role the confusion coefficient (relationship between predicted intermediate states under a leakage model from different key hypotheses), and that not only the SNR (even if low) is a key factor influencing the success rate. Additionally, our result show that we cannot conclude that the 4-bit S-boxes are generally significantly less resistant than the investigated 8-bit S-boxes. In particular, when considering inverse S-boxes we showed that 4-bit S-boxes may be more resistant.

For profiled attacks, we analyze several machine learning techniques to recover 4-bit and 8-bit intermediate states. Our results show that attacking 4-bit is somewhat easier than attacking 8-bit, with the difference mainly stemming from the varying number of classes in one or the other scenario. Still, that difference is not so apparent as one could imagine. Since we work with only a single feature and yet obtain a good accuracy in a number of test scenarios, we are confident (as our experiments also confirm) that adding more features will render classification algorithms even more powerful, which will result in an even higher accuracy. Finally, we did not consider any countermeasures for the considered lightweight algorithms, since the capacity for adding countermeasures is highly dependent on the environment (which we assume to be much more constrained than in the case of AES). However, our results show that a smart selection of S-boxes results in an inherent resilience (especially for 4-bit S-boxes). Moreover, we show that in case of highly restricted devices, in which

countermeasures on the whole cipher are not practically feasible, a designer may choose to only protect the weakest round (first round) in the cipher to increase the side-channel resistant until a certain limit.

Our work in [23] concentrates on how to improve SCA resilience of ciphers without imposing any extra cost. This is possible by considering the inherent resilience of ciphers. We particularly concentrate on block ciphers which utilize S-boxes and therefore study the resilience of S-boxes against side-channel attacks. When discussing how to improve side-channel resilience of a cipher, an obvious direction is to use various masking or hiding countermeasures. However, such schemes come with a cost, e.g. an increase in the area and/or reduction of the speed. When considering lightweight cryptography and various constrained environments, the situation becomes even more difficult due to numerous implementation restrictions. However, some options are possible like using S-boxes that are easier to mask or (more on a fundamental level), using S-boxes that possess higher inherent side-channel resilience. In [23] we investigate what properties should an S-box possess in order to be more resilient against side-channel attacks. Moreover, we find certain connections between those properties and cryptographic properties like nonlinearity and differential uniformity. Finally, to strengthen our theoretical findings, we give an extensive experimental validation of our results.

[64] Side-channel Analysis of Lightweight Ciphers: Current Status and Future Directions

[65] Side-channel Analysis of Lightweight Ciphers: Does Lightweight Equal Easy?

[10] Lightweight Ciphers and their Side-channel Resilience.

[23] Trade-Offs for S-Boxes: Cryptographic Properties and Side-Channel Resilience

[24] Do we need a holistic approach for the design of secure IoT systems? hal-01628683

7.1.5. New Advances on Side-channel Distinguishers

Participants: Axel Legay, Annelie Heuser.

[16] *Template Attack vs Bayes Classifier*

Side-channel attacks represent one of the most powerful category of attacks on cryptographic devices with profiled attacks in a prominent place as the most powerful among them. Indeed, for instance, template attack is a well-known real-world attack that is also the most powerful attack from the information theoretic perspective. On the other hand, machine learning techniques have proven their quality in a numerous applications where one is definitely side-channel analysis. As one could expect, most of the research concerning supervised machine learn- ing and side-channel analysis concentrated on more powerful machine learning techniques. Although valid from the practical perspective, such attacks often remain lacking from the more theoretical side. In this paper, we investigate several Bayes classifiers, which present simple supervised techniques that have significant similarities with the template attack. More specifically, our analysis aims to investigate what is the influence of the feature (in)dependence in datasets with different amount of noise and to offer further insight into the efficiency of machine learning for side-channel analysis.

[46] *Side-channel analysis and machine learning: A practical perspective* The field of side-channel analysis has made significant progress over time. Analyses are now used in practice in design companies as well as in test laboratories, and the security of products against side-channel attacks has significantly improved. However, there are still some remaining issues to be solved for analyses to be more effective. Side-channel analysis ac- tually consists of two steps, commonly referred to as identification and exploitation. The identification consists of understanding the leakage in order to set up a relevant attack. On the other hand, the exploitation consists of using the identified leakages to extract the secret key. In scenarios where the model is poorly known, it can be approximated in a profiling phase. There, machine learning techniques are gaining value. In this paper, we conduct extensive analysis of several machine learning techniques, showing the importance of proper parameter tuning and training. In contrast to what is perceived as common knowledge in unrestricted scenarios, we show that some machine learning techniques can significantly outperform template attack when properly used. We therefore stress that the traditional worst case security assessment of cryptographic implementations that includes mainly template attacks might not be accurate enough. Besides that, we present a new measure called the Data Confusion Factor that can be used to assess how well machine learning techniques will perform on a certain dataset.

[30] *Codes for Side-Channel Attacks and Protections*

This article revisits side-channel analysis from the standpoint of coding theory. On the one hand, the attacker is shown to apply an optimal decoding algorithm in order to recover the secret key from the analysis of the side-channel. On the other hand, the side-channel protections are presented as a coding problem where the information is mixed with randomness to weaken as much as possible the sensitive information leaked into the side-channel. Therefore, the field of side-channel analysis is viewed as a struggle between a coder and a decoder. In this paper, we focus on the main results obtained through this analysis. In terms of attacks, we discuss optimal strategy in various practical contexts, such as type of noise, dimensionality of the leakage and of the model, etc. Regarding countermeasures, we give a formal analysis of some masking schemes.

[38] *Climbing Down the Hierarchy: Hierarchical Classification for Machine Learning Side-Channel Attacks*

Machine learning techniques represent a powerful paradigm in side-channel analysis, but they come with a price. Selecting the appropriate algorithm as well as the parameters can sometimes be a difficult task. Nevertheless, the results obtained usually justify such an effort. However, a large part of those results use simplification of the data relation and in fact do not consider allthe available information. In this paper, we analyze the hierarchical relation between the data and propose a novel hierarchical classification approach for side-channel analysis. With this technique, we are able to introduce two new attacks for machine learning side-channel analysis: Hierarchical attack and Structured attack. Our results show that both attacks can outperform machine learning techniques using the traditional approach as well as the template attack regarding accuracy. To support our claims, we give extensive experimental results and discuss the necessary conditions to conduct such attacks.

[14] *Stochastic Collision Attack*

On the one hand, collision attacks have been introduced in the context of side-channel analysis for attackers who exploit repeated code with the same data without having any knowledge of the leakage model. On the other hand, stochastic attacks have been introduced to recover leakage models of internally processed intermediate secret variables. Both techniques have shown advantages and intrinsic limitations. Most collision attacks, for instance, fail in exploiting all the leakages (e.g., only a subset of matching samples are analyzed), whereas stochastic attacks cannot involve linear regression with the full basis (while the latter basis is the most informative one). In this paper, we present an innovative attacking approach, which combines the flavors of stochastic and collision attacks. Importantly, our attack is derived from the optimal distinguisher, which maximizes the success rate when the model is known. Notably, we develop an original closed-form expression, which shows many benefits by using the full algebraic description of the leakage model. Using simulated data, we show in the unprotected case that, for low noise, the stochastic collision attack is superior to the state of the art, whereas asymptotically and thus, for higher noise, it becomes equivalent to the correlation-enhanced collision attack. Our so-called stochastic collision attack is extended to the scenario where the implementation is protected by masking. In this case, our new stochastic collision attack is more efficient in all scenarios and, remarkably, tends to the optimal distinguisher. We confirm the practicability of the stochastic collision attack thanks to experiments against a public data set (DPA contest v4). Furthermore, we derive the stochastic collision attack in case of zero-offset leakage that occurs in protected hardware implementations and use simulated data for comparison. Eventually, we underline the capability of the new distinguisher to improve its efficiency when the attack multiplicity increases.

[15] *Optimal side-channel attacks for multivariate leakages and multiple models*

Side-channel attacks allow to extract secret keys from embedded systems like smartcards or smartphones. In practice, the side-channel signal is measured as a trace consisting of several samples. Also, several sensitive bits are manipulated in parallel, each leaking differently. Therefore, the informed attacker needs to devise side-channel distinguishers that can handle both multivariate leak-

ages and multiple models. In the state of the art, these two issues have two independent solutions: on the one hand, dimensionality reduction can cope with multivariate leakage; on the other hand, online stochastic approach can cope with multiple models. In this paper, we combine both solutions to derive closed-form expressions of the resulting optimal distinguisher in terms of matrix operations, in all situations where the model can be either profiled offline or regressed online. Optimality here means that the success rate is maximized for a given number of traces. We recover known results for uni- and bivariate models (including correlation power analysis) and investigate novel distinguishers for multiple models with more than two parameters. In addition, following ideas from the AsiaCrypt?2013 paper ?Behind the Scene of Side-Channel Attacks,? we provide fast computation algorithms in which the traces are accumulated prior to computing the distinguisher values.

[39] *Stochastic Side-Channel Leakage Analysis via Orthonormal Decomposition*

Side-channel attacks of maximal efficiency require an accurate knowledge of the leakage function. Template attacks have been introduced by Chari et al. at CHES 2002 to estimate the leakage function using available training data. Schindler et al. noticed at CHES 2005 that the complexity of profiling could be alleviated if the evaluator has some prior knowledge on the leakage function. The initial idea of Schindler is that an engineer can model the leakage from the structure of the circuit. However, for some thin CMOS technologies or some advanced countermeasures, the engineer intuition might not be sufficient. Therefore, inferring the leakage function based on profiling is still important. In the state-of-the-art, though, the profiling stage is conducted based on a linear regression in a non-orthonormal basis. This does not allow for an easy interpretation because the components are not independent. In this paper, we present a method to characterize the leakage based on a Walsh-Hadamard orthonormal basis with staggered degrees, which allows for direct interpretations in terms of bits interactions. A straightforward application is the characterization of a class of devices in order to understand their leakage structure. Such information is precious for designers and also for evaluators, who can devise attack bases relevantly.

[17] *On the optimality and practicability of mutual information analysis in some scenarios*

The best possible side-channel attack maximizes the success rate and would correspond to a maximum likelihood (ML) distinguisher if the leakage probabilities were totally known or accurately estimated in a profiling phase. When profiling is unavailable, however, it is not clear whether Mutual Information Analysis (MIA), Correlation Power Analysis (CPA), or Linear Regression Analysis (LRA) would be the most successful in a given scenario. In this paper, we show that MIA coincides with the maximum likelihood expression when leakage probabilities are replaced by online estimated probabilities. Moreover, we show that the calculation of MIA is lighter that the computation of the maximum likelihood. We then exhibit two case-studies where MIA outperforms CPA. One case is when the leakage model is known but the noise is not Gaussian. The second case is when the leakage model is partially unknown and the noise is Gaussian. In the latter scenario MIA is more efficient than LRA of any order.

[59] *On the Relevance of Feature Selection for Profiled Side-channel Attacks*

In the process of profiled side-channel analysis there is a number of steps one needs to make. One important step that is often conducted without a proper attention is selection of the points of interest (features) within the side-channel measurement trace. Most of the related work start with an assumption that the features are selected and various attacks are then considered and compared to find the best approach. In this paper, we concentrate on the feature selection step and show that if a proper selection is done, most of the attack techniques offer satisfactory results. We investigate how more advanced feature selection techniques stemming from the machine learning domain can be used to improve the side-channel attack efficiency. Our results show that the so-called Hybrid feature selection methods result in the best classification accuracy over a wide range of test scenarios and number of features selected.

[60] *Profiled SCA with a New Twist: Semi-supervised Learning*

Profiled side-channel attacks represent the most powerful category of side-channel attacks. In this context, the attacker gains ac- cess of a profiling device to build a precise model which is used to attack another device in the attacking phase. Mostly, it is assumed that the attacker has unlimited capabilities in the profiling phase, whereas the attacking phase is very restricted. We step away from this assumption and consider an attacker who is restricted in the profiling phase, while the attacking phase is less limited as in the traditional view. Clearly, in general, the attacker is not hindered to exchange any available knowledge between the profiling and attacking phase. Accordingly, we propose the concept of semi-supervised learning to side-channel analysis, in which the attacker uses the small amount of labeled measurements from the profiling phase as well as the unlabeled measurements from the attacking phase to build a more reliable model. Our results show that semi-supervised learning is beneficial in many scenarios and of particular interest when using template attack and its pooled version as side-channel attack techniques. Besides stating our results in varying scenarios, we discuss more general conclusions on semi-supervised learning for SCA that should help to transfer our observations to other settings in SCA.

7.1.6. Side-channel analysis on post-quantum cryptography

Participants: Axel Legay, Annelie Heuser, Tania Richmond, Martin Moreau.

In recent years, there has been a substantial amount of research on quantum computers ? machines that exploit quantum mechanical phenomena to solve mathematical problems that are difficult or intractable for conventional computers. If large-scale quantum computers are ever built, they will be able to break many of the public-key cryptosystems currently in use. This would seriously compromise the confidentiality and integrity of digital communications on the Internet and elsewhere. The goal of post-quantum cryptography (also called quantum-resistant cryptography) is to develop cryptographic systems that are secure against both quantum and classical computers, and can interoperate with existing communications protocols and networks. At present, there are several post-quantum cryptosystems that have been proposed: lattice-based, code-based, multivariate cryptosystems, hash-based signatures, and others. However, for most of these proposals, further research is needed in order to gain more confidence in their security and to improve their performance. Our interest lies in particular on the side-channel analysis and resistance of these post-quantum schemes. We first focus on code-based cryptography and then extend our analysis to find common vulnerabilities between different families of post-quantum crypto systems.

7.1.7. Binary Code Analysis: Formal Methods for Fault Injection Vulnerability Detection

Participants: Axel Legay, Thomas Given-Wilson, Annelie Heuser, Nisrine Jafri, Jean-Louis Lanet.

Formal methods such as model checking provide a powerful tool for checking the behaviour of a system. By checking the properties that define correct system behaviour, a system can be determined to be correct (or not).

Increasingly fault injection is being used as both a method to attack a system by a malicious attacker, and to evaluate the dependability of the system. By finding fault injection vulnerabilities in a system, the resistance to attacks or faults can be detected and subsequently addressed.

A process is presented that allows for the automated simulation of fault injections. This process proceeds by taking the executable binary for the system to be tested, and validating the properties that represent correct system behaviour using model checking. A fault is then injected into the executable binary to produce a mutant binary, and the mutant binary is model checked also. A different result to the validation of the executable binary in the checking of the mutant binary indicates a fault injection vulnerability.

This process has been automated with existing tools, allowing for easy checking of many different fault injection attacks and detection of fault injection vulnerabilities. This allows for the detection of fault injection vulnerabilities to be fully automated, and broad coverage of the system to be formally shown.

The work is implemented in the SimFi tool.

[56] (J; submitted) Fault injection has increasingly been used both to attack software applications, and to test system robustness. Detecting fault injection vulnerabilities has been approached with a variety of different but limited methods. This paper proposes an extension of a recently published general model checking based process to detect fault injection vulnerabilities in binaries. This new extension makes the general process scalable to real-world implementations which is demonstrated by detecting vulnerabilities in different cryptographic implementations.

7.1.8. Security at the hardware and software boundaries

Participants: Axel Legay, Jean-Louis Lanet, Ronan Lashermes, Kevin Bukasa, Hélène Le Bouder.

7.1.8.1. Side-channel attacks (SCA)

SCA exploit the reification of a computation through its physical dimensions (current consumption, EM emission, etc.). Focusing on Electromagnetic Analyses (EMA), such analyses have mostly been considered on low-end devices: smartcards and micro-controllers. In the wake of recent works, we analyze the effects of a modern micro architecture [31] on the efficiency of EMA (here Correlation Power Analysis and template attacks). We show that despite the difficulty to synchronize the measurements, the speed of the targeted core and the activity of other cores on the same chip can still be accommodated. Finally, we confirm that enabling the secure mode of TrustZone (a hardware-assisted software countermeasure) has no effect whatsoever on the EMA efficiency. Therefore, critical applications in TrustZone are not more secure than in the normal world with respect to EMA, in accordance with the fact that it is not a countermeasure against physical attacks. We hint that such techniques may be more common in the future to overcome the true difficulty with high-end devices: dealing with time precision (problem even worse with an OS or a virtual machine). Here again TrustZone or the activity of other cores have no incidence. But with these attacks, managing the big amount of data generated by our measures may prove to be the limiting factor, requiring better computing resources.

We investigate the way the compiler works and new attack paths have been discovered. In particular we demonstrated experimentally on an ARM7m the possibility to execute arbitrary code, generate buffer overflow even in presence of compiler assisted canary and ROP attacks. This raises a new challenge: any code fragment of an embedded program is sensitive to a fault attack. Thus an attacker increases the success rate of its attack while targeting a non sensitive part of the program for the injection. Then it becomes easy to extract security materials from the device. Then, the verification of the absence of a potential vulnerability must be checked on the whole program and not only on the cryptographic primitives. Thus the prevention analysis that was possible thanks to formal methods becomes unreachable with these new attack paths [40].

7.1.8.2. SCA based fuzzer

One of the main challenges during the development of system is to give a proof of evidence that its functionalities are correctly implemented and that no vulnerability remains. This objective is mostly achieved via testing techniques, which include software testing to check whether a system meets its functionalities, or security testing to express what should not happen. For the latter case, fuzzing is considered as first class citizen. It consists in exercising the system with (randomly) generated and eventually modified inputs in order to test its resistance. While fuzzing is definitively the fastest and the easiest way for testing applications, it suffers from severe limitations. Indeed, the precision of the model used for input generation: a random and/or simple model cannot reach all states and significant values. Moreover, a higher model precision can result in a combinatorial explosion of test cases.

We suggest a new approach [11] whose main ingredient is to combine timing attacks with fuzzing techniques. This new approach, allows not only reducing the test space explosion, but also to simplify the fuzzing process configuration. This new testing scenario is based on observing several executions of the system and by freezing some of its parameters in order to establish a partial order on their timing evaluation. The root of our technique is to exploit timing information to classify the input data into sub-domains according to the behavior observed for specific values of the parameters. Our approach is able to discover hidden unspecified commands that may trigger computations in the tested software. Due to the specific nature of the application (the domain of the parameters is the byte) and its programming model we can also retrieve the control flow graph of

the application. The limits of the approach have been identified, and it has been tested on two applications. Validation via a coverage tool has been established.

7.1.9. System Vulnerability Analysis

Participants: Jean-Louis Lanet, Abdelhal Mesbah, Razika Lounas, Chaharezd Yayaoui.

We present in this section our effort to detect and correct some misbehaviors encountered with some firmware. We start with an attack on a secure device, such that we are able to reverse a code while the ISA is unknown and the code itself is not available. Then, we propose a formal specification of the update process of a firmware which provides the guarantee that the updated program respects the semantics of the language. In a last aspect, we try to predict the ability of a program to be attacked thanks to a Machine Learning algorithm. We demonstrated in section 7.1.8 that a state exploration is useless until the whole program is examined, we demonstrated here that approximative solutions can deal with real live programs with an affordable response time.

7.1.9.1. Reverse engineering

We believe that an adversary can gain access to different assets of the system using a black box approach.This implies of course the absence of the source code, but also sometime the absence of the binary code (romized within the soc or micro-controller, no update mechanism, no jtag, no memory extraction, no read function, and so on). In that case, the first step consists in extracting the binary code from the system. The attacker is just allowed to load data. He has then to infer enough information on the system internals and then he should be able to gain access to the native layers. In [43], we demonstrate the advantage of a graphical representation of the data in the memory can help the reverse process thanks to the abstraction provided. Our graphical tool links all the objects with a relationship based on the presence of a pointer.

In a Java based secure element, a Java application is considered as data executed by the executed program (the virtual machine) by the native processor. We introduce a first weakness in the program that allows to read an instance as an array which violate the Java type system. This weakness allows us to dump a short part of the memory which contains the meta data on a set of arrays. Thanks to this information, we generate a mimicry attack by forging pointer illegally [41]. In turns, it open the possibility to read large part of the memory as element of a forged array. Then we succeed in characterizing the memory management algorithm [12]. At the end, we transform the initial problem of finding a vulnerability in the code of a device in a black box approach to a white box problem after de-assembling the binary code.

In another work [44], we studied the byte code verification process towards an unchecked code. We found that this verification is not complete and can be bypassed. The verifier checks the semantics of the Java Card byte code. This process is split in two parts. First, the verifier loads the methods' byte code and checks the package content. For the method segment, it checks that the control flow remain inside the methods, the jump destinations are correct and so on. Secondly, for each entry point and only for these, it controls the semantics and the type correctness of the code. This step is not performed for unreachable code, while the specification states that no unreachable code should remain in the file. However, during our analysis we discovered that the verifier does some verification on the semantics of the unreachable code. Then, thanks to a fault attack (the return byte code is noped) we diverted the control flow into this unchecked area were we stored our ill-typed code leading to the execution of an aggressive shell code which in turn dumped the native layers of the card giving access to the secret key material in plain text.

7.1.9.2. Safe system update mechanism

Dynamic Software Updating (DSU) consists in updating running programs on the fly without any downtime. This feature is interesting in critical applications that must run continuously. Because updates may lead to security breaches, the question of their correctness is raised. Formal methods are a rigorous means to ensure the correctness required by applications using DSU. We propose [13] a formal verification of correctness of DSU in a Java-based embedded system. Our approach is based on three steps. First, a formal interpretation of the semantics of update operations to ensure type safety of the update. Secondly, we rely on a functional representation of byte code, the predicate transformation calculus, and a functional model of the update

mechanism to ensure the behavioral correctness of the updated programs. It is based on the use of Hoare predicate transformation to derive a specification of an updated byte code. In the last step, we use the functional representation to model the safe update point detection mechanism. This mechanism guarantees that none of the updated method active methods are active. This property is called activeness safety. We propose a functional specification that allows to derive proof obligations that guarantee the safety of the mechanism.

7.1.9.3. Prediction of system divergence

Fault attack represents one of the serious threats against embedded system security. The result of the fault injection could lead to a mutation of the code in such a way that it becomes hostile or execute a unwanted sequence of code as we demonstrated in 7.1.8 . Any successful attack may reveal a secret information stored in the card or grant an undesired authorization. We propose a methodology [5] to recognize, during the development step, the sensitive patterns to the fault attack. It is based on the concepts from text categorization and machine learning. In fact, in this method we represented the patterns using opcodes n-grams as features and we evaluated different machine learning classifiers.

In the first experiment, we evaluated all the combination of n-gram size (for n=2, n=3 and n=4), number of features using GR method to select 100, 200, ..., 500 and 1000 relevant n-grams, n-gram weighting (Term Frequency (TF), Term Frequency Inverse Document Frequency (TFIDF) and binary representations), and five classification algorithms (Naive Bayes network (NB), Decision Tree (DT), Support Vector Machine (SVM), and the boosted version of these two lasts (BDT and BSVM)) to determine the best setting. We used accuracy measure to evaluate performance of the classifiers. In addition to accuracy, we used F1, TP rate and FP rate measures to evaluate how the algorithms classified the dangerous patterns. In the first experiment, we noted that 2-gram outperformed others. Nearly 2-gram, TFIDF, 1000 features with boosted algorithm outperformed the other settings. The F1 results have shown that the classifiers are more accurate at classifying examples of the class of non dangerous pattern compared to other classes. We suggest that this might be due to the imbalance of our data set. In the second experiment, we investigated the imbalance problem. We applied SMOTE and NCR resampling techniques to overcome this class imbalance problem. We found that the outperforming setting in the resampled data set was St_{270} also with BSVM classifier. Resampled data set improves accuracy of the smallest class and keeps the accuracy of other classes.

The experimental results indicated that the resampling techniques improved the accuracy of the classifiers. In addition, our proposed method reduces the execution time of sensitive patterns classification in comparison to the mutant generator tool micro seconds instead of hours.

7.2. Results for Axis 2: Malware analysis

The detection of malicious programs is a fundamental step to be able to guarantee system security. Programs that exhibit malicious behavior, or *malware*, are commonly used in all sort of cyberattacks. They can be used to gain remote access on a system, spy on its users, exfiltrate and modify data, execute denial of services attacks, etc.

Significant efforts are being undertaken by software and data companies and researchers to protect systems, locate infections, and reverse damage inflicted by malware. Our contribution to malware analysis include the following fields:

7.2.1. Malware Detection

Participants: Axel Legay, Fabrizio Biondi, Olivier Decourbe, Mike Enescu, Thomas Given-Wilson, Annelie Heuser, Jean-Louis Lanet, Jean Quilbeuf, Alexander Zhdanov, Olivier Zendra.

Given a file or data stream, the malware detection problem consists of understanding if the file or data stream contain traces of malicious behavior. For binary executable files in particular, this requires extracting a signature of the file, so it can be compared against signatures of known clean and malicious files to determine whether the file is malicious. Binary file signatures can be divided in *syntactic* and *semantic*.

Syntactic signatures are based on properties of the file itself, like its length, hash, number and entropy of the executable and data sections, and so on. While syntactic signatures are computationally cheap to extract from binaries, it is also easy for malware creators to deploy *obfuscation* techniques that change the file's syntactic properties, hence widely mutating the signature and preventing its use for malware detection.

Semantic signatures instead are based on the binary's behavior and interactions with the system, hence are more effective at characterizing malicious files. However, they are more expensive to extract, requiring behavioral analysis and reverse-engineering of the binary. Since behavior is much harder to change than syntactic properties, against these signatures obfuscation is used to harden the file against reverse-engineering and preventing the analysis of the behavior, instead of changing it directly.

In both cases, *malware deofbuscation* is necessary to extract signatures containing actuable information that can be used to characterize the binaries as clean or malicious. Once the signatures are available, *malware classification* techniques, usually based on machine learning, are used to automatically determine whether binaries are clean or malicious starting from their signatures. Our contributions on these fields are described in the next sections.

7.2.2. Malware Deobfuscation

Participants: Axel Legay, Fabrizio Biondi, Olivier Decourbe, Mike Enescu, Thomas Given-Wilson, Annelie Heuser, Nisrine Jafri, Jean-Louis Lanet, Jean Quilbeuf.

Given a file (usually a portable executable binary or a document supporting script macros), deobfuscation refers to the preparation of the file for the purposes of further analysis. Obfuscation techniques are specifically developed by malware creators to hinder detection reverse engineering of malicious behavior. Some of these techniques include:

Packing Packing refers to the transformation of the malware code in a compressed version to be dynamically decompressed into memory and executed from there at runtime. Packing techniques are particularly effective against static analysis, since it is very difficult to determine statically the content of the unpacked memory to be executed, particularly if packing is used multiple times. The compressed code can also be encrypted, with the key being generated in a different part of the code and used by the unpacking procedure, or even transmitted remotely from a command and control (C&C) server.

Control Flow Flattening This technique aims to hinder the reconstruction of the control flow of the malware. The malware's operation are divided into basic blocks, and a dispatcher function is created that calls the blocks in the correct order to execute the malicious behavior. Each block after its execution returns control to the dispatcher, so the control flow is flattened to two levels: the dispatcher above and all the basic blocks below.

To prevent reverse engineering of the dispatcher, it is often implemented with a cryptographic hash function. A more advanced variant of this techniques embed a full virtual machine with a randomly generated instruction set, a virtual program counted, and a virtual stack in the code, and uses the machine's interpreter as the dispatcher.

Virtualization is a very effective technique to prevent reverse engineering. To contrast it, we are implementing state-of-the-art devirtualization algorithms in `angr` , allowing it to detect and ignore the virtual machine code and retrieving the obfuscated program logic. Again, we plan to contribute our improvements to the main `angr` branch, thus helping the whole security community fighting virtualized malware.

Opaque Constants and Conditionals Reversing packing and control flow flattening techniques requires understanding of the constants and conditionals in the program, hence many techniques are deployed to obfuscate them and make them unreadable by reverse engineering techniques. Such techniques are used e.g. to obfuscate the decryption keys of packed encrypted code and the conditionals in the control flow.

We have proven the efficiency of dynamic synthesis in retrieving opaque constant and conditionals, compared to the state-of-the-art approach of using SMT (Satisfiability Modulo Theories) solvers, when the input space of the opaque function is small enough. We are developing techniques based on fragmenting and analyzing by brute force the input space of opaque conditionals, and SMT constraints in general, to be integrated in SMT solvers to improve their effectiveness.

7.2.3. Malware Classification

Participants: Axel Legay, Fabrizio Biondi, Olivier Decourbe, Mike Enescu, Thomas Given-Wilson, Annelie Heuser, Nisrine Jafri, Jean-Louis Lanet, Jean Quilbeuf.

Once malicious behavior has been located, it is essential to be able to classify the malware in its specific family to know how to disinfect the system and reverse the damage inflicted on it.

While it is rare to find an actually previously unknown malware, morphic techniques are employed by malware creators to ensure that different generations of the same malware behave differently enough than it is hard to recognize them as belonging to the same family. In particular, techniques based on the syntax of the program fails against morphic malware, since syntax can be easily changed.

To this end, semantic signatures are used to classify malware in the appropriate family. Semantic signatures capture the malware's behavior, and are thus resistant to morphic and differentiation techniques that modify the malware's syntactic signatures. We are investigating semantic signatures based on the program's System Call Dependency Graph (SCDG), which have been proven to be effective and compact enough to be used in practice. SCDGs are often extracted using a technique based on pushdown automata that is ineffective against obfuscated code; instead, we are applying concolic analysis via the `angr` engine to improve speed and coverage of the extraction.

Once a semantic signature has been extracted, it has to be compared against large database of known signatures representing the various malware families to classify it. The most efficient way to obtain this is to use a supervised machine learning classifier. In this approach, the classifier is trained with a large sample of signatures malware annotated with the appropriate information about the malware families, so that it can learn to quickly and automatically classify signatures in the appropriate family. Our work on machine learning classification focuses on using SCDGs as signatures. Since SCDGs are graphs, we are investigating and adapting algorithms for the machine learning classification of graphs, usually based on measures of shared subgraphs between different graphs. One of our analysis techniques relies on common subgraph extraction, with the idea that a malicious behavior characteristic of a malware family will yield a set of common subgraphs. Another approach relies on the Weisfeiler-Lehman graph kernel which uses the presence of nodes and their neighborhoods pattern to evaluate similarity between graphs. The presence or not of a given pattern becomes a feature in a subsequent machine learning analysis through random forest or SVM.

In malware detection and classification, it is fundamental to have a false positive rate (i.e. rate of cleanware classified as malware) approaching zero, otherwise the classification system will classify hundred or thousands of cleanware files as malware, making it useless in practice. To decrease the false positive rate, the classifier is also trained with a large and representative database of cleanware, so that it can discriminate between signatures of cleanware and malware with a minimal false positive rate. We use a large database of malware and cleanware to train our classifier, thus guaranteeing a high detection rate with a small false positive rate.

We have put in place a platform for malware analysis, using dedicated hardware provided by Cisco. This platform is now fully operational and receives a daily feed of suspicious binaries for analysis. Furthermore, we developed tools for maintaining our datasets of cleanware and malware binaries, run existing syntactic analysis on them. Our toolchain is able to extract SCDGs from malwares and cleanwares and apply our classification techniques on the SCDGs.

7.2.4. Botnet Trojan Detection

Participants: Axel Legay, Fabrizio Biondi, Vesselin Bontchev, Thomas Given-Wilson, Jean Quilbeuf, Olivier Decourbe, Najah Ben Said.

Botnet trojans are a class of malware that opens a backdoor in a system and waits from further instructions from a C&C server, and possibly replicates itself somehow. A large group of systems infected by such malware is known as a botnet, and can be used by the botnet's controller to distribute spam emails (possibly carrying other malware) and perform distributed denial-of-service (DDoS) attacks. In a DDoS attack, all the systems in the botnet flood a single target with requests amounting to gigabytes or even terabytes of traffic. The target is not able to handle such traffic or to discriminate malicious request from legitimate ones, failing to provide its service.

Detecting and correctly classify botnet trojans in transit is a necessary step to be able to stop their infection. We applied our semantic classification approach on a particular family of malware, the Mirai botnet. With these experiments, we were able to confirm that the classification based on SCDG extraction and common subgraphs mining has a very low false positive rate and a high detection rate. Furthermore, our approach proved to be more accurate than detection based on syntactic signatures, without increasing the number of false positives.

7.2.5. Modular Automated Syntactic Signature Extraction (MASSE)

Participants: Axel Legay, Fabrizio Biondi, Olivier Zendra, Alexander Zhdanov, Bruno Lebon, François Déchelle.

Malware detection techniques based on syntactic signatures (or "rules") are commonly used in antivirus since their low computational cost allows them to be used on scan the files handled by the system without excessively slowing down the system. Semantic analysis techniques are relatively expensive to use, and would slow down a system significantly if used for on-access malware detection. Hence, it is common in antivirus company to use advanced semantic techniques like the SCDG-based ones we develop to detect and analyze known and unknown malware samples, and then to manually write a syntactic rule for the detection of such samples that is uploaded to the client machines.

The MASSE projects aims at providing an open-source, self-contained architecture to deploy this on a given system, company, or infrastructure, without needing to give access to the structure's data to third parties. The architecture is composed of a server executing the computationally-expensive semantic analysis, and of a number of lightweight clients performing inexpensive syntactic analysis on the client's systems. The MASSE server automatically analyzes unknown or suspicious files passing on the clients, detects the malicious ones, synthesizes syntactic signatures for them, and updates the signature databases of the clients, keeping them protected.

The MASSE server exploits modular malware analysis, supporting malware analysis modules using dynamic, static, or hybrid analysis; extracting syntactic, semantic, or hybrid signatures; using signature-based or anomaly-based detection; and any other technique the user desires, thanks to its open source malware analysis APIs. MASSE also implements pseudonymization of the signature databases, preventing an attacker to learn precisely the syntactic signatures in case some of the clients are compromised.

7.2.6. Malware IDS

Participants: Jean-Louis Lanet, Aurélien Palisse, Colas Le Guernic.

7.2.6.1. An efficient IDS for malware detection

Ransomware is a type of malware that prevents legitimate users from accessing their machine or files and demands a payment for restoring the functionalities of the infected computer. There are two classes of ransomware: the *simple lockers*, which block the usage of the computer, and *cryptors*, that encrypt files on the computer. In the case of encryption-based ransomware, the user data can only be restored with the secret key(s) used during the attack if the key is provided by the attacker.

Detecting a malware can use two options:

- The system knows the features of the malware. Features can be structural information: n-gram or graph isomorphism, or behavioral information: APIs call or system calls. Exact pattern matching algorithm or approximative algorithm (Machine learning) can be used. This approach is known as signature based and can only detect known patterns.

- The system knows its correct behavior. Any deviation of this model leads to the detection of hostile programs. This approach can detect any new attack, it does not rely on a model of the bad behavior but on the model of the correct behavior. This approach is also known as IDS (Intrusion Detection System).

In [45], [34] we apply this technique to detect malware at run time (EPS: End Point Solution). Our first solution is based on the dynamic analysis of the data transformation by the program. We propose to monitor file activity. Since it has already been proven a valid approach in terms of detection, our main goal in is to show that a good detection rate can be achieved with little to no impact on system performances. To this end, we limit our monitoring to a minimum. In order to reduce the impact on detection with a low rate of false positive, we use the chi-square goodness-of-fit test instead of Shannon entropy (*i.e.*, sensitive to compressed chunks of data). We also achieve system completeness and fine granularity by monitoring the whole file system for all userland threads. In order to evaluate our prototype implementation, Data Aware Defense (DaD), under realistic conditions, we used the bare-metal analysis platform of the LHS, Malware - O - Matic (MoM), and ran it on a large and heterogeneous (compared to the literature) live ransomware collection. We used *de facto* industry standard benchmarks to get a pertinent and reproducible assessment of the performance penalties. A second model of the correct behavior with better results has been developed (patent pending).

Our countermeasure is efficient and can be deployed on Windows 7/10 machines with a reasonable performance hit, with an average delay of 12 μs per write operation on disk, a few hundred times smaller than previous approaches. Our extensive experiments show that the more sophisticated ransomware already use mimicry attacks. However we successfully detect 99.37 % of the samples with at most 70 MB lost per sample's threads in 90% of cases and less than 7 MB in 70% of cases. Its speed and low negative rate makes it a good candidate as a first line of defense. Once a thread is deemed malicious, instead of blocking disk accesses, other more costly metrics can be used to improve the false positive rate without impacting performance, since it would not be computed for all other threads.

7.2.7. Papers

This section gathers papers that are results common to all sections above pertaining to Axis 2.

[51] (C) The largest DDoS attacks in history have been executed by devices controlled by the Mirai botnet trojan. To prevent Mirai from spreading, this paper presents and evaluates techniques to classify binary samples as Mirai based on their syntactic and semantic properties. Syntactic malware detection is shown to have a good detection rate and no false positives, but to be very easy to circumvent. Semantic malware detection is resistant to simple obfuscation and has better detection rate than syntactic detection, while keeping false positives to zero. This paper demonstrates these results, and concludes by showing how to combine syntactic and semantic analysis techniques for the detection of Mirai.

[19] (C) We present the MASSE architecture, a YARA-based open source client-server malware detection platform. MASSE includes highly effective automated syntactic malware detection rule generation for the clients based on a server-side modular malware detection system. Multiple techniques are used to make MASSE effective at detecting malware while keeping it from disrupting users and hindering reverse-engineering of its malware analysis by malware creators.

[4] (J) Control flow obfuscation techniques can be used to hinder software reverse-engineering. Symbolic analysis can counteract these techniques, but only if they can analyze obfuscated conditional statements. We evaluate the use of dynamic synthesis to complement symbolic analysis in the analysis of obfuscated conditionals. We test this approach on the taint-analysis-resistant Mixed Boolean Arithmetics (MBA) obfuscation method that is commonly used to obfuscate and randomly diversify statements. 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 targeted algebraic simplification can greatly reduce the analysis time. We show that synthesis-based deobfuscation is more effective than current SMT-based deobfuscation algorithms, thus proposing a synthesis-based attacker model to complement existing attacker models.

7.3. Results for Axis 3: Building a secure network stack

7.3.1. Privacy-Preserving Abuse Detection in Future Decentralised Online Social Networks

Participants: Jeffrey Burdges, Alvaro Garcia Recuero, Christian Grothoff.

Future online social networks need to not only protect sensitive data of their users, but also protect them from abusive behavior coming from malicious participants in the network. We investigated the use of supervised learning techniques to detect abusive behavior and describe privacy-preserving protocols to compute the feature set required by abuse classification algorithms in a secure and privacy-preserving way. While our method is not yet fully resilient against a strong adaptive adversary, our evaluation suggests that it will be useful to detect abusive behavior with a minimal impact on privacy.

Our results show how to combine local knowledge with private set intersection and union cardinality protocols (with masking of BLS signature to protect identity of signers/subscribers) to privately derive feature values from users in OSNs. Given an adaptive adversary that would be able to manipulate most features we propose in our supervised learning approach, it is surprising that with just three features resistant to adversarial manipulation, the algorithms still provide useful classifications.

This work was originally presented at DPM 2016 [63] and expanded upon in Álvaro García-Recuero's PhD thesis [1].

7.3.2. Fog of Trust

Participants: Jeffrey Burdges, Christian Grothoff.

The Web of Trust (WoT) used traditionally used by tools for private communication such as PGP is used to to validate individual links between participants. Using the WoT, however, leaks meta data, such that users must opt-in for it – exposing themselves to risks of privacy loss. We proposed a new method, the Fog of Trust (FoT), which uses the privacy-preserving set intersection cardinality protocol originally used in our work on abuse detection in online social networks, to support this critical step of public key verification via collaboration. In the FoT, the social relationships — which are used to verify public keys – remain hidden. This allows keys to be verified via trusted intermediaries that were established beforehand, without the need to verify each individual new contact using Trustwords. Consequently, FoT will can the same functionality as the WoT without its drawbacks to privacy.

7.3.3. Cell tower privacy

Participants: Christian Grothoff, Neal Walfield.

Context-aware applications are programs that are able to improve their performance by adapting to the current conditions, which include the user's behavior, networking conditions, and charging opportunities. In many cases, the user's location is an excellent predictor of the context. Thus, by predicting the user's future location, we can predict the future conditions. In this work, we developed techniques to identify and predict the user's location over the next 24 hours with a minimum median accuracy of 82results include our observation that cell phones sample the towers in their vicinity, which makes cell towers as-is inappropriate for use as landmarks. Motivated by this observation, we developed two techniques for processing the cell tower traces so that landmarks more closely correspond to locations, and cell tower transitions more closely correspond to user movement. We developed a prediction engine, which is based on simple sampling distributions of the form $f(t, c)$, where t is the predicted tower, and c is a set of conditions. The conditions that we considered include the time of the day, the day of the week, the current regime, and the current tower. Our family of algorithms, called TomorrowToday, achieves 89% prediction precision across all prediction trials for predictions 30 minutes in the future. This decreases slowly for predictions further in the future, and levels off for predictions approximately 4 hours in the future, at which point we achieve 82% prediction precision across all prediction trials up to 24 hours in the future. This represents a significant improvement over NextPlace, a well-cited prediction algorithm based on non-linear time series, which achieves appropriately 80% prediction precision (self reported) for predictions 30 minutes in the future, but, unlike our predictors, which try all prediction attempts, NextPlace only attempts 7% of the prediction trials on our data set [67].

7.3.4. Taler protocol improvements

Participants: Jeffrey Burdges, Florian Dold, Christian Grothoff, Marcello Stanisci.

We started modeling the Taler protocol in the framework of Provable Security, precisely defining the formal meaning of income transparency, fairness, anonymity and unforgeablity as security games. The resulting definitions and security proofs allow a more precise statement of the security of Taler in relation to the security assumptions that are being made.

The implementation of the wallet module now supports the full Taler protocol, including the refresh operation for highly efficient and privacy-preserving change.

In addition to improving the stability of the implementation of all Taler components, we added new features to the protocol that (1) allow refunds from merchants without violating privacy and (2) allow merchants to do "customer tipping", which transfers money from merchants directly to customers' wallets as a reward for doing actions on their website.

7.3.5. Mix Networking

Participants: Jeffrey Burdges, Christian Grothoff.

We have begun implementing our ratcheting scheme for providing hybrid post-quantum and forward security to the Sphinx mix network packet format. We also began collaborating with the Panoramix project and LEAP to help resolve numerous practical challenges to deploying a mix network. We shall speak about this ongoing work at the Chaos Computer Club's annual congress 34c3 in December 2017.

7.4. Other research results

7.4.1. Privacy and Security: Information-Theoretical Quantification of Security Properties

Participants: Axel Legay, Fabrizio Biondi, Olivier Zendra, Thomas Given-Wilson, Annelie Heuser, Sean Sedwards, Jean Quilbeuf, Mike Enescu.

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 information 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 information by combining this information with their prior knowledge of the system.

Armed with the produced output of the system, the attacker tries to infer information about the secret information that produced the output. The quantitative analysis we consider defines and computes how much information the attacker can expect to infer (typically measured in bits). This expected leakage of bits is the information leakage of the system. This is computed by symbolically exploring the code to be analyzed, and using the symbolic constraints accumulated over the output together with a model counting algorithm to quantify the leakage.

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 that don't it is imperative to be able to distinguish between the systems leaking very small amounts of secret information and systems leaking a significant amount of secret information, since only the latter are considered to pose a security vulnerability to the system.

While quantitative leakage computation is a powerful technique to detect security vulnerabilities, computing the leakage of complex programs written in low-level languages is a hard and computationally intensive task. The most common language for low-level implementation of security protocols is C, due to its efficiency, hence much of the effort in developing tools to detect vulnerabilities in source code focus on C. Recently, we have improved the state of the art in leakage quantification from C programs by proposing the usage of approximated model counting instead of precise model counting. We have shown how the approximation can improve the efficiency of leakage quantification by orders of magnitude against a logarithmic decrease in the precision of the result, often producing the same result as precise model counters much faster, and often being able to analyze cases where precise model counters would have failed. We demonstrated this technique by providing the first quantitative leakage analysis of the C code of the Heartbleed bug, showing that our technique can detect the bug in the code.

A different but equally interesting approach is followed by our new HyLEak tool. HyLeak is also able to analyze a system and compute its information leakage, i.e. the amount of information that an observer would gain by about the value of system's secret by observing its output. Contrarily to other techniques, HyLeak can analyze randomized systems, and correctly distinguish between the randomness injected in the system and the uncertainty on the secret value. This allows HyLeak to be used both on systems with explicit randomization and systems that depend on stochastic properties, like cyber-physical systems.

HyLeak uses static code analysis to divide the system to be analyzed in components. For each component, HyLeak evaluates whether it is more convenient to analyze the component using precise or statistical analysis. Each component is analyzed with the most appropriate strategy, and then the results for all components are combined together and information leakage is estimated.

The hybrid approach provides better results than both the precise and the statistical ones in terms of computation time and precision of the result. Also, it bridges the gap between cheap but imprecise statistical techniques and precise but expensive formal techniques, allowing the user to control the required precision of the result according to the computation time they have available. We evaluated HyLeak against QUAIL's precise approach and the statiatical approach implemented in the LeakWatch tool, showing that HyLeak outperforms them both. HyLeak is open source and available at https://project.inria.fr/hyleak/

Applied to shared-key cryptosystems, the information-theoretical approach allows precise reasoning about the information leakage of any secret information in the system including, the key, and the message. Recent work on max-equivocation has generalised perfect secrecy and shown the maximum achievable theoretic bounds for the security of the key and message. Achieving these theoretic maximal bounds has been proven to be achievable by Apollonian Cell Encoders (ACEs). ACEs not only allow the maximum security possible in a shared-key cryptosystem, but also allow for infinite key reuse when the key has less entropy than the message. Further, ACEs are straightforward to construct and have a compact representation making them feasible to use in practice.

Another application is to use information leakage to reason about leakage through shared resources, representing various side-channel attacks. Developmens here allow for the formalising of the leakage model through shared resources, and quantifying how significant the leakage can be. This improves on the state-of-the-art that uses only qualified leakage, and so can be precise about how much is leakage through a shared resource. Such quantification of leakage allows for scheduling of the shared resource to exploit this information to minimise leakage. Such minimisation of leakage allows for scheduling and utilisation of resources that would fail a simple quanlified test, providing solutions when prior state-of-the-art would claim impossibility. Further, a reasoned trade-off can be made between acceptable leakage and utility of the shared resource, allowing solutions that are acceptable even if not perfect.

[53] (C; submitted) Preserving privacy of private communication against an attacker is a fundamental concern of computer science security. Unconditional encryption considers the case where an attacker has unlimited computational power, hence no complexity result can be relied upon for encryption. Optimality criteria are defined for the best possible encryption over a general collection of entropy measures. This paper introduces Apollonian cell encoders, a class of shared-key cryptosystems that are proven to be universally optimal. In addition to the highest possible security for the message,

Apollonian cell encoders prove to have perfect secrecy on their key allowing unlimited key reuse. Conditions for the existence of Apollonian cell encoders are presented, as well as a constructive proof. Further, a compact representation of Apollonian cell encoders is presented, allowing for practical implementation.

[18] (C) High-security processes have to load confidential information into shared resources as part of their operation. This confidential information may be leaked (directly or indirectly) to low-security processes via the shared resource. This paper considers leakage from high-security to low-security processes from the perspective of scheduling. The workflow model is here extended to support preemption, security levels, and leakage. Formalization of leakage properties is then built upon this extended model, allowing formal reasoning about the security of schedulers. Several heuristics are presented in the form of compositional preprocessors and postprocessors as part of a more general scheduling approach. The effectiveness of such heuristics are evaluated experimentally, showing them to achieve significantly better schedulability than the state of the art. Modeling of leakage from cache attacks is presented as a case study.

[52] (C) Quantitative information flow measurement techniques have been proven to be successful in detecting leakage of confidential information from programs. Modern approaches are based on formal methods, relying on program analysis to produce a SAT formula representing the program's behavior, and model counting to measure the possible information flow. However, while program analysis scales to large codebases like the OpenSSL project, the formulas produced are too complex for analysis with precise model counting. In this paper we use the approximate model counter ApproxMC2 to quantify information flow. We show that ApproxMC2 is able to provide a large performance increase for a very small loss of precision, allowing the analysis of SAT formulas produced from complex code. We call the resulting technique ApproxFlow and test it on a large set of benchmarks against the state of the art. Finally, we show that ApproxFlow can evaluate the leakage incurred by the Heartbleed OpenSSL bug, contrarily to the state of the art.

[20] (C) We present HyLeak, a tool for reasoning about the quantity of information leakage in programs. The tool takes as input the source code of a program and analyzes it to estimate the amount of leaked information measured by mutual information. The leakage estimation is mainly based on a hybrid method that combines precise program analysis with statistical analysis using stochastic program simulation. This way, the tool combines the best of both symbolic and randomized techniques to provide more accurate estimates with cheaper analysis, in comparison with the previous tools using one of the analysis methods alone. HyLeak is publicly available and is able to evaluate the information leakage of randomized programs, even when the secret domain is large. We demonstrate with examples that HyLeaks has the best performance among the tools that are able to analyze randomized programs with similarly high precision of estimates.

[54] (J; submitted) Analysis of a probabilistic system often requires to learn the joint probability distribution of its random variables. The computation of the exact distribution is usually an exhaustive precise analysis on all executions of the system. To avoid the high computational cost of such an exhaustive search, statistical analysis has been studied to efficiently obtain approximate estimates by analyzing only a small but representative subset of the system's behavior. In this paper we propose a hybrid statistical estimation method that combines precise and statistical analyses to estimate mutual information, Shannon entropy, and conditional entropy, together with their confidence intervals. We show how to combine the analyses on different components of the system with different accuracy to obtain an estimate for the whole system. The new method performs weighted statistical analysis with different sample sizes over different components and dynamically finds their optimal sample sizes. Moreover it can reduce sample sizes by using prior knowledge about systems and a new abstraction-then-sampling technique based on qualitative analysis. To apply the method to the source code of a system, we show how to decompose the code into components and to determine the analysis method for each component by overviewing the implementation of those techniques in HyLeak tool. We demonstrate with case studies that the new method outperforms the state of the art in quantifying information leakage.

7.4.2. Security for therapeutical environments

Participants: Axel Legay, Olivier Zendra, Thomas Given-Wilson, Sean Sedwards.

This work is done in the context of the ACANTO EU project. We aim at helping develop robotic assistants to aid mobility of mobility-impaired and elderly adults. These robotic assistants provide a variety of support to their users, including: navigational assistance, social networking, social activity planning, therapeutic regime support, and diagnostic support. In Tamis, we focus on navigational assistance and social activities, as together they yield an interesting challenge in human robot interaction. The goal is to help groups of users navigate in a potentially busy dynamic environment, while also maintaining social group cohesion.

A robotic assistant has been developed before in the DALi project, acting selfishly to ensure the safe navigation of a single user. This was achieved by using the social force model and statistical model checking in a reactive planner that frequently replanned and made immediate navigational suggestions to the user. The key operational loop of this solution was to: observe the environment, model the agents in the environment in the social force model, give safety constraints for the user, and then use statistical model checking to find the optimal next move for the user.

Generalising to groups of users poses several significant difficulties. Computationally, the challenge is exponential in the number of users, considering all their possible navigational choices. Incomplete information is normal, since sensors are distributed between robotic assistants and the environment, and communication may fail, leading to different robots having different knowledge of the environment. Maintaining group cohesion is non-trivial, since group composition and position are dynamic and, unlike swarm robotics, no group member can be abandoned. Frequent replanning is necessary since there is minimal control over the users' actions, which may include ignoring the advise of the robotic assistant

The solution we designed is to abstract away from individual users in favour of groups. This refines the prior solution for a single user. Sensor information is used to obtain traces that provide behavioural information about users and pedestrians in the environment. These traces are clustered into groups that capture both location and motion behaviour. The groups are used as the social particles in the social force model, with parameters adjusted to account for group dynamics. Statistical model checking is used to find the optimal next move for the group containing the user, and the navigation for the optimal next move is displayed to the user. The effectiveness of the group abstraction mechanisms use in this refined algorithm are validated on the BIWI walking pedestrians dataset. This shows they operate correctly and effectively, even improving over human annotations, on real world data of pedestrians in a chaotic environment.

[27] (C) 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 the ACANTO project is developing a robotic assistant that allows its users to engage in therapeutic group social activities, building on work done in the DALi project. Key components of the ACANTO technology are social networking and group motion planning, both of which entail the sharing and broadcasting of information. Given that the system may also make use of medical records, it is clear that the issues of security, privacy, and trust are of supreme importance to ACANTO.

[58] (C; submitted) The ACANTO project is developing robotic assistants to aid the mobility and recovery of mobility-impaired and older adults. One key feature of the project's robotic assistants is aiding with navigation in chaotic environments. Prior work has solved this for a single user with a single robot, however for therapeutic outcomes ACANTO supports social groups and group activities. Thus these robotic assistants must be able to efficiently support groups of users walking together. This requires an efficient navigation solution that can handle large numbers of users, maintain (de-facto) group cohesion despite unpredictable behaviours, and operate rapidly on embedded devices. We address these challenges by: using sensor information to develop behavioural traces, clustering traces to determine groups, modeling the groups using the social force model, and finding an optimal navigation solution using statistical model checking. The new components of this solution are validated on the ETH Zürich dataset of pedestrians in an open environment.

7.4.3. Mobile air pollution sensor platform for smart-cities

Participant: Laurent Morin.

This work is organized and coordinated by the Chaire “mobilité dans une ville durable” and financed by the Foundation of Rennes 1 (https://fondation.univ-rennes1.fr/)

The purpose of this work is to design and experiment a mobile pollution sensor platform for Smart-Cities in Rennes.

The platform is integrated in the project ROAD (Rennes Open Access to Data) proposing to development of mobile systems operating the collection and the management of open data in Rennes for a future development of a smart-city. The collaboration is part of an ecosystem developed by the Chair “mobilité dans une ville durable” via the production of multiple experimentations in the city.

In the ROAD project context, the air quality in the city has been identified as one of the major challenge. Air quality improvement can only be achieved with a citizen and political full cooperation and involvement. This experimentation aims at providing an end-to-end urban platform that extends current practices in air quality measurements and allows citizens and policy makers to obtain the data and make informed decisions.

The mobile air pollution sensor platform for smart-cities proposes a innovative IoT architecture introducing the deployment of a small set of advanced and cost-effective sensors around a balanced high-performance/low-power compute unit inside a mobile agent in the city. The compute unit will have to provide the necessary computation power needed to produce advanced analysis and the security management on-site (integrity, authentication, ...).

The mobile sensor platform developments partially started in July 2017, and accelerated in October for a real deployment in buses in 2018. During this period, the core system of the platform was designed, adapted, and partially implemented to offer an operational prototype. This year lead to the design of a suitcase containing a self-sufficient measurement system: a main compute unit, its power supply and power management, and a set of satellite pollution sensors. This achievement was disseminated to the Rennes ecosystem (Rennes Atalante, Rennes Métropole, Inria) through the participation to several meetings and exhibitions.

TEA Project-Team

7. New Results

7.1. ADFG: Affine data-flow graphs scheduler synthesis

Participants: Loïc Besnard, Thierry Gautier, Alexandre Honorat, Jean-Pierre Talpin, Hai Nam Tran.

We consider with ADFG (Affine DataFlow Graph) the synthesis of periodic scheduling parameters for real-time systems modeled as ultimately cyclo-static dataflow (UCSDF) graphs [14]. This synthesis aims for a trade-off between throughput maximization and total buffer size minimization. The synthesizer inputs are: a UCSDF graph which describes tasks by their Worst Case Execution Time (WCET), and directed buffers connecting tasks by their data production and consumption rates; the number of processors in the target system and the real-time scheduling synthesis algorithm to be used. The outputs are the synthesized scheduling parameters: the tasks periods, offsets, processor bindings and priorities, and the buffers initial marking and maximum sizes.

ADFG was originally the implementation of Adnan Bouakaz's work [0]. However the tool had not been packaged yet to be easily installed and used. Moreover, code refactoring led to improve the theory, and to add new features. Firstly, more accurate bounds and Integer Linear Programming (ILP) formulations have been used. Besides, dataflow graphs do not need to be weakly connected for EDF policy on multiprocessor systems. The new implementation also avoids to use a fixed parameter for some multiprocessor partitioning algorithms, now an optional strategy enables to compute it. Finally implementation has been adapted to standard technologies to be more easily installed and used. As the synthesizer evolved a lot, new evaluations have been made. Moreover, many scheduled examples have been simulated with Cheddar [0], which provides pertinent metrics to analyze the scheduling efficiency.

ADFG is being extended to investigate and solve the scheduling problem of dataflow programs on many-core architectures. These architectures have distinctive traits requiring significant changes to classical multiprocessor scheduling theory. There is a high number of contention points introduced by novel memory architectures and new interconnect types such as Network-on-Chip. Two solutions are proposed and implemented in ADFG: contention-aware and contention-free scheduling synthesis. We either take into account the contention and synthesize a contention-aware schedule or find a schedule that results in no contention.

7.2. Formal Semantics of Behavior Specifications in the Architecture Analysis and Design Language Standard

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

The Architecture Analysis and Design Language (AADL) is a standard proposed by SAE to express architecture specifications and share knowledge between the different stakeholders about the system being designed. To support unambiguous reasoning, formal verification, high-fidelity simulation of architecture specifications in a model-based AADL design workflow, we have defined formal semantics for the behavior specification of the AADL. These semantics rely on the structure of automata present in the standard already, yet provide tagged, trace semantics framework to establish formal relations between (synchronous, timed, asynchronous) usages or interpretations of behavior [17]. We define the model of computation and communication of a behavior specification by the synchronous, timed or asynchronous traces of automata with variables. These constrained automata are derived from *polychronous automata* defined within the polychronous model of computation and communication [11].

[0] *Real-Time Scheduling of Dataflow Graphs.* A. Bouakaz. PhD Thesis, University of Rennes 1, 2013.
[0] *The Cheddar project: a GPL real-time scheduling analyzer.* http://beru.univ-brest.fr/~singhoff/cheddar/

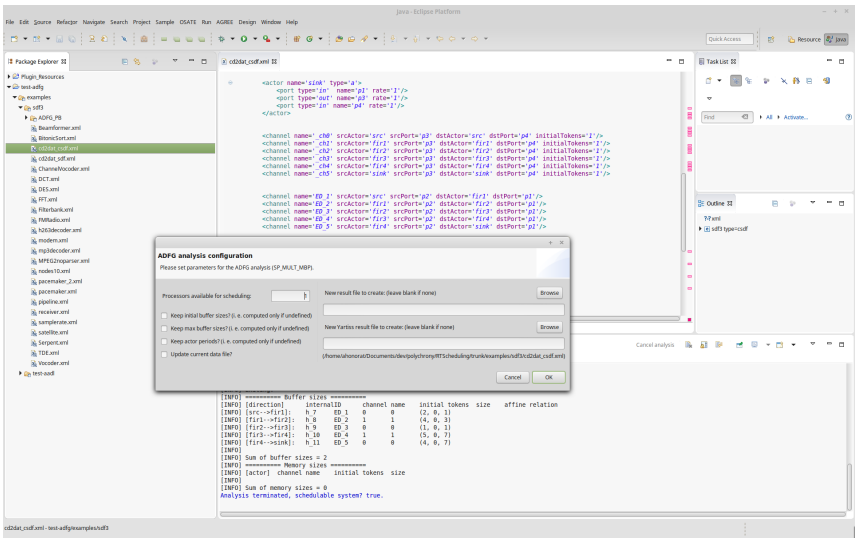

Figure 1. ADFG under Eclipse

States of a behavior annex transition system can be either observable from the outside (*initial*, *final* or *complete* states), that is states in which the execution of the component is paused or stopped and its outputs are available; or non observable execution states, that is internal states. We thus define two kinds of steps in the transition system: *small steps*, that is non-observable steps from or to an internal state; and *big steps*, that is observable steps from a *complete* state to another, through a number of small steps). The semantics of the AADL considers the observable states of the automaton. The set of states S_A of automaton A (used to interpret the behavior annex) thus only contains states corresponding to these observable states and the set of transitions T_A big-step transitions from an observable state to another (by opposition with small-step transitions from or to an execution state). The action language of the behavior annex defines actions performed during transitions. Actions associated with transitions are action blocks that are built from basic actions and a minimal set of control structures (sequences, sets, conditionals and loops). Typically, a behavior action sequence is represented by concatenating the transition systems of its elements; a behavior action set is represented by composing the transition systems of its elements.

The polychronous model of computation had been used previously as semantic model for systems described in the core AADL standard. This translation of AADL specifications into the polychronous model now takes into account the behavior specifications. The import of AADL behavior annexes (AADL-BA) to the polychronous model relies on polychronous automata and on small steps/big steps semantics. Small steps may be viewed as an implicit oversampling of the big steps. To express such implicit upsampling, a model of *Signal-thread* has been introduced in Polychrony (refer to Section “New trends and developments in Polychrony”). In that context, the translation of a behavior annex associated with an AADL thread consists mainly in the production of the corresponding Signal automaton, which is declared as a Signal-thread, and the definition of the environment required for this Signal-thread. In particular, the signal *complete-thread* is defined so that it will occur when the next state of the automaton is a *complete* state (the control will return to the scheduler): in other words, it specifies the end of a sequence of small steps.

A specific difficulty in the translation of AADL-BA is the translation of the action language, which is related to the general problem of the translation of a sequential language to a dataflow one. First, in AADL-BA actions, a given variable may be assigned several times in a sequence (for example, $x = a + b; x = x + a$). Thus an AADL-BA action has to be transformed into a SSA (static single assignment) form ($x_0 = a + b; x = x_0 + a$

in the previous example). Another possible problem is the translation of AADL-BA loop structures (for, while, do until). In our case, this is solved, again, by considering them as Signal-threads: the *dispatch-thread* event is defined by the upperbound of the clocks of the inputs of the loop and the *complete-thread* event defines the termination of the loop.

7.3. New trends and developments in Polychrony

Participants: Loïc Besnard, Thierry Gautier.

The synchronous modeling paradigm provides strong correctness guarantees for embedded system design while requiring minimal environmental assumptions. In most related frameworks, global execution correctness is achieved by ensuring the insensitivity of (logical) time in the program from (real) time in the environment. This property, called endochrony, can be statically checked, making it fast to ensure design correctness. Unfortunately, it is not preserved by composition, which makes it difficult to exploit with component-based design concepts in mind.

It has been shown that compositionality can be achieved by weakening the objective of endochrony: a weakly endochronous system is a deterministic system that can perform independent computations and communications in any order as long as this does not alter its global state. Moreover, the non-blocking composition of weakly endochronous processes is isochronous, which means that the synchronous and asynchronous compositions of weakly endochronous processes accept the same behaviors. Unfortunately, testing weak endochrony needs state-space exploration, which is very costly in the general case. Then, a particular case of weak endochrony, called polyendochrony, was defined, which allows static checking thanks to the existing clock calculus. The clock hierarchy of a polyendochronous system may have several trees, with synchronization relations between clocks placed in different trees, but the clock expressions of the clock system must be such that there is no clock expression (especially, no root clock expression) defined by symmetric difference: root clocks cannot refer to absence. In other words, the clock system must be in disjunctive form [9].

We have now implemented code generation for polyendochronous systems in Polychrony. This generation reuses techniques of distributed code generation, with rendez-vous management for synchronization constraints on clocks which are not placed in the same tree of clocks. For such a synchronization constraint $c_1 = c_2$, nodes *send* and *receive* are added in the graph, associated with clocks c_1 and c_2: for c_1, *send*(c_1) is followed by *receive*(c_2), followed itself by all the other nodes associated with clock c_1; and symmetrically for c_2. Then the subgraphs corresponding respectively to the trees where c_1 and c_2 are placed are separated, as if they were distributed on different processors. In this way, nodes *send* and *receive* become respectively outputs and inputs (both for c_1 and c_2) of the subgraphs. Finally, a communication library (MPI) is used for simulation. The following restriction is considered in the current implementation: the roots of the trees of c_1 and c_2 must be free variables.

We have also considered another extension related to clocks, again for making code generation possible for more programs than it was the case before. A characteristic of the Signal language is that it allows to specify programs which have internal accelerations with respect to their inputs and outputs. However, the constraint that implemented programs, for which code was generated, should be endochronous, restricted more or less these programs to have one single such acceleration (or clock upsampling). To abstract from this restriction, we have defined a model of so-called *Signal-thread*, that helps to confine such accelerations, and thus to generate code for programs with multiple clock upsampling. A Signal-thread is a Signal process with internal implicit upsampling; it has a *dispatch-thread* input event and a *complete-thread* output event; its outputs are delayed compared with its inputs. As the Signal-thread represents an upsampling, the *step* (see [1]) of the corresponding generated code is a loop. Such Signal-threads may be considered as a pragmatic way to implement *clock domains*.

7.4. Modular verification of cyber-physical systems using contract theory

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

The primary goal of our project, in collaboration with Mitsubishi Electronics Research Centre Europe (MERCE), is to ensure correctness-by-design in realistic cyber-physical systems, i.e., systems that mix software and hardware in a physical environment, e.g., Mitsubishi factory automation lines or water-plant factory. To achieve that, we develop a verification methodology based on decomposition into components enhanced with contract reasoning.

The work of A. Platzer on Differential Dynamic Logic ($d\mathcal{L}$) holds our attention [0]. This a formalism built on the Dynamic Logic of V. Pratt augmented with the possibility of expressing Ordinary Differential Equations (ODEs), which are the usual way to model physical behaviors in physics. Combined with the ability of Dynamic Logic to specify and verify hybrid programs, $d\mathcal{L}$ is particularly fit model cyber-physical systems. The proof system associated with the logic is implemented into the theorem prover KeYmaera X. Aimed toward automatisation, it is a promising tool to spread formal methods into industry.

We have defined a syntactic parallel composition operator in $d\mathcal{L}$ which enjoys associativity and commutativity [15]. Commutativity allows to compose component in every possible order. Associativity is mandatory to modularly design a system; it allow to upgrade a system by adding new components. We have then characterized the conditions under which we can derive automatically a proof of the contract of our composition of two components, given the proof of the contract for each components. Theses theoretical results have been exemplified with an example of a cruise-controller entirely proved within the interactive theorem prover KeYmaera X.

The study of the cruise-controller example and of a water-tank system highlights some limitations of our approach. We can not handle retro-action and we have to compose in parallel components which have to be sequenced, e.g. a sensor and a computer. We have overcomed theses limitations by introducing a sequential composition operator which enjoys associativity and distributivity over the parallel composition operator. We believe it is a first step toward a composition algebra in $d\mathcal{L}$. This operator also satisfy the property that we can automatically derive a proof of the contract of our composition of two components, given the proof of the contract for each components, but under some relaxed conditions. We believe it is the first step toward a composition algebra.

Thanks to these results, a wide variety of systems are now possible to modularly design in $d\mathcal{L}$. To validate our approach, we are currently working on the implementation of our parallel composition operator as a tactic in KeYmaera X.

To challenge our ideas, we are working in the proof of a realistic cyber-physical system, a power-train system used in automotive. We plan to use it as a basis to test abstraction mechanisms to ultimately allow mix between top-down and bottom-up design.

7.5. Parametric verification of time synchronization protocols

Participants: Ocan Sankur, Jean-Pierre Talpin.

In the context of the associate-team COMPOSITE, we addressed the verification one of the apparently simplest services in any loosely-coupled distributed system : the time service. In many instances of such systems, traffic and power grids, banking and transaction networks, the accuracy and reliability of this service are critical.

In the instance of sensor networks, it is of particular interest to verify the robustness of such protocols to variations caused by the environment. Lake of power, varying temperatures, imperfect hardware, are sources of local drifts and jitters in time measurement that require self-calibration and fault-tolerance to reach distributed consensus. FTSP, the flooding time synchronization protocol, provides fault-tolerance and enables time synchronization.

In [16], we introduce an environment abstraction technique and an incremental model checking technique to prove that FTSP eventually elects a leader for any network topology and configuration (anonymized identifiers), up to a diameter $N = 7$ (with synchronous communications) and $N = 5$ (desynchronized communications), resulting in significant improvements over previous results.

[0] *Differential Dynamic Logic for Hybrid Systems*, André Platzer, http://symbolaris.com/logic/dL.html

7.6. Modular analysis and verification of system libraries

Participants: Jean-Joseph Marty, Jean-Pierre Talpin.

We are starting to develop a new perspective on the active topic of information flow control (IFC). We plan to adapt current investigations to tagged multi-core architecture, including software (virtual machines) and hardware (the Risc V processor) experiments and applications. All this work is based on the previous experience about verified Unikernel programming on low resources processors such as the Arduino (Marty's Master internship). We will define formally relations between processes and blocks of code inside a concurrent environment. This line of work will be investigated for both embedded IoT applications and cloud computing. By working with IFC at processor level and system level, we will enforce strong security foundation and focus on constraint solving analysed software.

ASPI Team

5. New Results

5.1. Central limit theorem for adaptive multilevel splitting

Participants: Frédéric Cérou, Arnaud Guyader, Mathias Rousset.

See 3.2 , and 4.2 .

This is a collaboration with Bernard Delyon (université de Rennes 1).

Fleming–Viot type particle systems represent a classical way to approximate the distribution of a Markov process with killing, given that it is still alive at a final deterministic time. In this context, each particle evolves independently according to the law of the underlying Markov process until its killing, and then branches instantaneously on another randomly chosen particle. While the consistency of this algorithm in the large population limit has been recently studied in several articles, our purpose here is to prove central limit theorems under very general assumptions. For this, we only suppose that the particle system does not explode in finite time, and that the jump and killing times have atomless distributions. In particular, this includes the case of elliptic diffusions with hard killing.

5.2. Adaptive multilevel splitting for Monte Carlo particle transport

Participant: Mathias Rousset.

See 3.2 , and 4.2 .

Simulation of neutron transport with Monte Carlo methods is a central issue in order to assess the aging of french nucelar plants.

In [49], we propose an alternative version of the AMS (adaptive multilevel splitting) algortihm, adapted for the first time to the field of particle tranport. Within this context, it can be used to build an unbiased estimator of any quantity associated with particle tracks, such as flux, reaction rates or even non–Boltzmann tallies. Furthermore, the effciency of the AMS algorithm is shown not to be very sensitive to variations of its input parameters, which makes it capable of significant variance reduction without requiring extended user effort.

5.3. Weak overdamped limit theorem for Langevin processes

Participant: Mathias Rousset.

This is a collaboration with Pierre-André Zitt (université Paris Est Marne-la-Vallée).

The Langevin stochastic process is the main model used in molecular dynamics simulation, for instance for the simulation of reactive trajectories of bio-chemical systems with rare event techniques.

In [21], we prove convergence in distribution of Langevin processes in the overdamped diffusion asymptotics. The proof relies on the classical perturbed test function (or corrector) method, which is used both to show tightness in path space, and to identify the extracted limit with a martingale problem. The result holds assuming the continuity of the gradient of the potential energy, and a mild control of the initial kinetic energy.

5.4. Particle–Kalman filter for structural health monitoring

Participant: Frédéric Cérou.

This is a joint work with EPI I4S (Inria Rennes–Bretagne Atlantique).

Standard filtering techniques for structural parameter estimation assume that the input force either is known exactly or can be replicated using a known white Gaussian model. Unfortunately for structures subjected to seismic excitation, the input time history is unknown and also no previously known representative model is available. This invalidates the aforementioned idealization. To identify seismic induced damage in such structures using filtering techniques, a novel algorithm is proposed to estimate the force as additional state in parallel to the system parameters. Two concurrent filters are employed for parameters and force respectively. For the parameters, interacting particle–Kalman filter is employed targeting systems with correlated noise. Alongside a second filter is employed to estimate the seismic force acting on the structure. The proposal is numerically validated on a sixteen degrees–of–freedom mass–spring–damper system. The estimation results confirm the applicability of the proposed algorithm.

In another work, the same approach has been used for varying system parameters with correlated state and observation noise. The idea is to nest a bank of linear KFs (Kalman filters) for state estimation within a PF (particle filter) environment that estimates the parameters. This facilitates employing relatively less expensive linear KF for linear state estimation problem while costly PF is employed only for parameter estimation. Additionally, the proposed algorithm also takes care of those systems for which system and measurement noises are not uncorrelated as it is commonly idealized in standard filtering algorithms. As an example, for mechanical systems under ambient vibration it happens when acceleration response is considered as measurement. Thus the process and measurement noise in these system descriptions are obviously correlated. For this, an improved description for the Kalman gain is developed. Further, to enhance the consistency of particle filtering based parameter estimation involving high dimensional parameter space, a new temporal evolution strategy for the particles is defined. This strategy aims at restricting the solution from diverging (up to the point of no return) because of an isolated event of infeasible estimation which is very much likely especially when dealing with high dimensional parameter space.

5.5. Reduced modeling of unknown trajectories

Participant: Patrick Héas.

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

In [12], we deal with model order reduction of parametrical dynamical systems. We consider the specific setup where the distribution of the system's trajectories is unknown but the following two sources of information are available: *(i)* some "rough" prior knowledge on the system's realisations, and *(ii)* a set of "incomplete" observations of the system's trajectories. We propose a Bayesian methodological framework to build reduced–order models (ROMs) by exploiting these two sources of information.

We emphasise that complementing the prior knowledge with the collected data provably enhances the knowledge of the distribution of the system's trajectories. We then propose an implementation of the proposed methodology based on Monte Carlo methods. In this context, we show that standard ROM learning techniques, such as proper orthogonal decomposition (POD) or dynamic mode decomposition (DMD), can be revisited and recast within the probabilistic framework considered in this work. We illustrate the performance of the proposed approach by numerical results obtained for a standard geophysical model.

5.6. Model reduction from partial observations

Participant: Patrick Héas.

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

In [11], we deal with model-order reduction of parametric partial differential equations (PPDE). More specifically, we consider the problem of finding a good approximation subspace of the solution manifold of the PPDE when only partial information on the latter is available. We assume that two sources of information are available: *i)* a "rough" prior knowledge, taking the form of a manifold containing the target solution manifold, and *ii)* partial linear measurements of the solutions of the PPDE (the term partial refers

to the fact that observation operator cannot be inverted). We provide and study several tools to derive good approximation subspaces from these two sources of information. We first identify the best worst-case performance achievable in this setup and propose simple procedures to approximate the corresponding optimal approximation subspace. We then provide, in a simplified setup, a theoretical analysis relating the achievable reduction performance to the choice of the observation operator and the prior knowledge available on the solution manifold.

5.7. Low–rank dynamic mode decomposition: optimal solution in polynomial time

Participant: Patrick Héas.

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

The works [15] and [41] study the linear approximation of high–dimensional dynamical systems using low-rank dynamic mode decomposition (DMD). Searching this approximation in a data–driven approach can be formalised as attempting to solve a low-rank constrained optimisation problem. This problem is non–convex and state–of–the–art algorithms are all sub–optimal. We show that there exists a closed-form solution, which can be computed in polynomial time, and characterises the ℓ_2–norm of the optimal approximation error. The theoretical results serve to design low–complexity algorithms building reduced models from the optimal solution, based on singular value decomposition or low–rank DMD. The algorithms are evaluated by numerical simulations using synthetic and physical data benchmarks.

5.8. Optimal kernel–based dynamic mode decomposition

Participant: Patrick Héas.

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

The state–of–the–art algorithm known as kernel-based dynamic mode decomposition (K–DMD) provides a sub–optimal solution to the problem of reduced modeling of a dynamical system based on a finite approximation of the Koopman operator. It relies on crude approximations and on restrictive assumptions. The purpose of the work in [20] is to propose a kernel–based algorithm solving exactly this low–rank approximation problem in a general setting.

5.9. Non parametric state–space model for missing–data imputation

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

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

Missing data are present in many environmental data–sets and this work aims at developing a general method for imputing them. State–space models (SSM) have already extensively been used in this framework. The basic idea consists in introducing the true environmental process, which we aim at reconstructing, as a latent process and model the data available at neighboring sites in space and/or time conditionally to this latent process. A key input of SSMs is a stochastic model which describes the temporal evolution of the environmental process of interest. In many applications, the dynamic is complex and can hardly be described using a tractable parametric model. Here we investigate a data-driven method where the dynamical model is learned using a non-parametric approach and historical observations of the environmental process of interest. From a statistical point of view, we will address various aspects related to SSMs in a non–parametric framework. First we will discuss the estimation of the filtering and smoothing distributions, that is the distribution of the latent space given the observations, using sequential Monte Carlo approaches in conjunction with local linear regression. Then, a more difficult and original question consists in building a non–parametric estimate of the dynamics which takes into account the measurement errors which are present in historical data. We will propose an EM–like algorithm where the historical data are corrected recursively. The methodology will be illustrated and validated on an univariate toy example.

I4S Project-Team

7. New Results

7.1. Outdoor InfraRed Thermography

7.1.1. Joint thermal and electromagnetic diagnostics

Participants: Nicolas Le Touz, Jean Dumoulin.

In this study, we present an inversion approach to detect and localize inclusions in thick walls under natural solicitations. The approach is based on a preliminary analysis of surface temperature field evolution with time (for instance acquired by infrared thermography); subsequently, this analysis is improved by taking advantage of a priori information provided by ground penetrating radar reconstruction of the structure under investigation. In this way, it is possible to improve the accuracy of the images achievable with the standalone thermal reconstruction method in the case of quasiperiodic natural excitation. [19]

7.1.2. Long term monitoring of transport infrastructures: from deployment to standardization

Participants: Antoine Crinière, Jean Dumoulin, Laurent Mevel.

Long term monitoring of transport infrastructures by infrared thermography has been studied and tested on different structures. A first standalone infrared system architecture developed is presented and discussed. Results obtained with such system on different Civil Engineering structures are presented. Some data processing approaches and inverse thermal model for data analysis are introduced and discussed. Lessons learned from experiments carried out in outdoor with such system are listed and analyzed. Then, a new generation of infrared system architecture is proposed. Finally, conclusions and perspectives are addressed.[29], [46]

7.1.3. Infrared data reconstruction and calibration for long term monitoring

Participants: Thibaud Toullier, Jean Dumoulin, Laurent Mevel.

This study focuses on the evaluation and improvement of thermal instrumentation solutions for long-term monitoring of next-generation transport infrastructure. A test site was equipped with thermocouples and an infrared thermography system coupled with monitoring of environmental parameters. A method of spatial reconstruction of infrared images is presented. Measurement data acquired on site and then post-processed are analysed over time. A conclusion on the results achieved and prospects are proposed [48]

7.2. Data management of Smart territories and cities

Participants: Antoine Crinière, Jean Dumoulin.

Highly instrumented Smart-cities, which are now a common urban policies, are facing problems of management and storage of a large volume of data coming from an increasing number of sources. This study presents a data compression method by predictive coding of spatially correlated multi-source data based on reference selection and prediction by Kriging [47]

7.3. Smarts roads and R5G

7.3.1. Energy exchange modelization and infrared monitoring for hybrid pavement structure

Participants: Nicolas Le Touz, Thibaud Toullier, Jean Dumoulin.

In those studies, we evaluate by numerical modelling the energy inputs that could occur in a hybrid pavement structure with a semi-transparent or opaque wearing course bonded to a porous base layer, the seat of a heat transfer fluid circulation. The digital studies conducted propose a coupled resolution of various thermal phenomena: diffusion/convection in the case of an opaque surface drainage pavement, and diffusion/convection/radiation for a pavement with a semi-transparent surface. Coupled equation systems are solved numerically using the finite element method. This model was developed directly on a Matlab kernel. In a second time, laboratory experiments on small specimen were carried out and the surface temperature was monitored by infrared thermography. Results obtained are analyzed and performances of the numerical model for real scale outdoor application are discussed .[35], [34]

7.3.2. Phase change materials characterization

Participant: Jean Dumoulin.

In a costs reduction and comfort requirements context, the use of phase change materials (PCM) is a sustainable and economical answer. For transportation infrastructures and winter maintenance, they avoid ice occurrence or snow accumulation. Their characteristics, and more specifically, the solid to liquid phase transition temperature and enthalpy, are usually obtained through DSC. Raman spectroscopy can bring answers and information on their microstructures. The liquid to solid phase change was investigated on three PCM, a paraffin, formic acid and diluted formic acid. A comparison made on freezing temperature obtained through DSC, Raman spectroscopy associated with chemiometrics indicated a consistency between the methods. Raman spectroscopy coupled with multivariate data analysis allowed the identification of an additional specificity in the freezing process of the paraffin. All methods provided results consistent between each other, although some differences between literature and experimental freezing temperatures of the considered PCM were observed in all cases. [20], [53]

7.4. Methods for building performance assessment

7.4.1. Building performance assessment

Participants: Jordan Brouns, Alexandre Nassiopoulos.

Two additive thermal sources are generally not simultaneously distinguishable from the only observation of their effect on the heat balance. However, there are cases where information about the variation regularity of these sources is known. This is typically the case of convective internal gains in the building, for which the use scenarios create discontinuous inputs while heat gains relating to the air leakage are regular in time. In the present paper, we introduce a method aiming to distinguish heat sources using this a priori knowledge about their dynamics. We provide numerical and experimental evidence that the method succeeds in separating/distinguishing these kind of sources. This method could be applied to the identification of the occupancy rate for measurement and verification plans or smart home systems such as learning thermostats. [16]

7.5. System identification

7.5.1. Variance estimation of modal parameters from subspace-based system identification

Participants: Michael Doehler, Laurent Mevel.

This work has been carried out in collaboration with Palle Andersen.

Subspace-based system identification allows the accurate estimation of the modal parameters (natural frequencies, damping ratios, mode shapes) from output-only measurements, amongst others with data-driven methods like the Unweighted Principal Component (UPC) algorithm. Due to unknown excitation, measurement noise and finite measurements, all modal parameter estimates are inherently afflicted by uncertainty. The information on their uncertainty is most relevant to assess the quality of the modal parameter estimates, or when comparing modal parameters from different datasets. A method for variance estimation is presented for the variance computation of modal parameters for the UPC subspace algorithm. Developing the sensitivities of the modal parameters with respect to the output covariances, the uncertainty is propagated from the measurements to the modal parameters from UPC. The resulting variance expressions are easy to evaluate and computationally tractable when using an efficient implementation. In a second step, the uncertainty information of the stabilization diagram is used to extract appropriately weighted global mode estimates and their variance. The method is applied to experimental data from the Z24 Bridge [30].

7.5.2. *Bayesian parameter estimation for parameter varying systems using interacting Kalman filters*

Participants: Antoine Crinière, Laurent Mevel, Jean Dumoulin, Subhamoy Sen.

This work is in collaboration with F. Cerou of ASPI team at Inria.

Standard filtering techniques for structural parameter estimation assume that the input force either is known exactly or can be replicated using a known white Gaussian model. Unfortunately for structures subjected to seismic excitation, the input time history is unknown and also no previously known representative model is available. A novel algorithm is proposed to estimate the force as additional state in parallel to the system parameters. Two concurrent filters are employed for parameters and force respectively, mixing interacting Particle Kalman filter and another filter employed to estimate the seismic force acting on the structure [38], [49].

7.5.3. *From structurally independent local LTI models to LPV model*

Participant: Qinghua Zhang.

This work on linear parameter varying (LPV) system identification has been carried out in collaboration with Lennart Ljung (Linköping University, Sweden).

The local approach to LPV system identification consists in interpolating individually estimated 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, individually estimated local state-space LTI models do not contain sufficient information for determining similarity transformations making them coherent. Nevertheless, it is possible to estimate these similarity transformations from input-output data under appropriate excitation conditions [21].

7.5.4. *Stability of the Kalman filter for output error systems*

Participant: Qinghua Zhang.

The stability of the Kalman filter is classically ensured by the uniform complete controllability regarding the process noise and the uniform complete observability of linear time varying systems. Recently we have studied the stability of the Kalman filter for output error (OE) systems, in which the process noise is totally absent. In this case the classical stability analysis assuming the controllability regarding the process noise is thus not applicable. Our first efforts were focused on continuous time systems, whereas discrete time systems have been studied since last year. It is shown in this work that the uniform complete observability is sufficient to ensure the stability of the Kalman filter applied to time varying OE systems, regardless of the stability of the OE systems [22].

7.5.5. *Reduced-order interval-observer design for dynamic systems with time-invariant uncertainty*

Participant: Qinghua Zhang.

This work on interval-based state estimation has been carried out in collaboration with Vicenç Puig's team (Universitat Politècnica de Catalunya, Spain). The reported work addresses in particular the design of reduced-order interval-observers for dynamic systems with both time-invariant and time varying uncertainties. Because of the limitations of the set-based approach and the wrapping effect to deal with interval-observers, the trajectory-based interval-observer approach is used with an appropriate observer gain. Due to difficulties to satisfy the conditions for selecting a suitable gain to guarantee the positivity of the resulting observer, a reduced-order observer is designed to increase the degree of freedom when selecting the observer gain and to reduce the computational complexity. Simulation examples illustrates the effectiveness of the proposed approach [37].

7.5.6. Parameter uncertainties quantification for finite element based subspace fitting approaches

Participants: Guillaume Gautier, Laurent Mevel, Michael Doehler.

This work has been carried out in collaboration with Jean-Mathieu Mencik and Roger Serra (INSA Centre Val de Loire).

Recently, a subspace fitting approach has been proposed for vibration-based finite element model updating. The approach makes use of subspace-based system identification, where the extended observability matrix is estimated from vibration measurements. Finite element model updating is performed by correlating the model-based observability matrix with the estimated one. However, estimates from vibration measurements are inherently exposed to uncertainty. A covariance estimation procedure for the updated model parameters is proposed, which propagates the data-related covariance to the updated model parameters by considering a first-order sensitivity analysis. In particular, this propagation is performed through each iteration step of the updating minimization problem, by taking into account the covariance between the updated parameters and the data-related quantities. Simulated vibration signals and experimental data of a beam validate the method [18].

7.6. Damage diagnosis

7.6.1. Damage detection by perturbation analysis and additive change detection theory

Participants: Michael Doehler, Laurent Mevel, Qinghua Zhang.

The monitoring of mechanical systems aims at detecting damages at an early stage, in general by using output-only vibration measurements under ambient excitation. In this paper, a method is proposed for the detection and isolation of small changes in the physical parameters of a linear mechanical system. Based on a recent work where the multiplicative change detection problem is transformed to an additive one by means of perturbation analysis, changes in the eigenvalues and eigenvectors of the mechanical system are considered in the first step. In a second step, these changes are related to physical parameters of the mechanical system. Finally, another transformation further simplifies the detection and isolation problem into the framework of a linear regression subject to additive white Gaussian noises, leading to a numerically efficient solution of the considered problems. A numerical example of a simulated mechanical structure is reported for damage detection and localization [31].

7.6.2. Damage localization using the statistical subspace damage localization method

Participants: Michael Doehler, Laurent Mevel, Saeid Allahdadian.

This work is happening during a thesis in collaboration with C. Ventura at UBC, Vancouver.

In this paper the statistical subspace damage localization (SSDL) method is employed in localizing the damage in a real structure, namely the Yellow frame. The SSDL method is developed for real testing conditions and tested in two damage configurations. It was demonstrated that the SSDL method can localize the damage robustly in the Yellow frame for simple and multiple distinct damage scenarios using the analytical modal parameters. The method is described and its effectiveness is demonstrated [24].

7.6.3. Stochastic Subspace-Based Damage Detection with Uncertainty in the Reference Null Space

Participants: Michael Doehler, Laurent Mevel, Eva Viefhues.

This work is happening during a thesis in collaboration with F. Hille at BAM, Berlin.

This paper deals with uncertainty considerations in damage diagnosis using the stochastic subspace-based damage detection technique. With this method, a model is estimated from data in a (healthy) reference state and confronted to measurement data from the possibly damaged state in a hypothesis test. Previously, only the uncertainty related to the measurement data was considered in this test, whereas the uncertainty in the estimation of the reference model has not been considered. We derive a new test framework, which takes into account both the uncertainties in the estimation of the reference model as well as the uncertainties related to the measurement data. Perturbation theory is applied to obtain the relevant covariances. In a numerical study the effect of the new computation is shown, when the reference model is estimated with different accuracies, and the performance of the hypothesis tests is evaluated for small damages. Using the derived covariance scheme increases the probability of detection when the reference model estimate is subject to high uncertainty, leading to a more reliable test [41].

7.6.4. Statistical damage localization with stochastic load vectors

Participants: Md Delwar Hossain Bhuyan, Michael Doehler, Laurent Mevel, Guillaume Gautier.

This work is in collaboration with F. Schoefs and Y. Lecieux, GEM, Nantes.

The Stochastic Dynamic Damage Locating Vector (SDDLV) method is a damage localization method based on both a Finite Element (FE) model of the structure and modal parameters estimated from measurements in the damage and reference states of the system. A vector is obtained in the null space of the changes in the transfer matrix from both states and then applied as a load vector to the model. The damage location is related to this stress where it is close to zero. An important theoretical limitation was that the number of modes used in the computation could not be higher than the number of sensors located on the structure. In this paper, the SDDLV method has been extended with a joint statistical approach for multiple mode sets, overcoming this restriction on the number of modes. Another problem is that the performance of the method can change considerably depending of the Laplace variable where the transfer function is evaluated. Particular attention is given to this choice and how to optimize it. The new approach is validated in numerical simulations and on experimental data. From these results, it can be seen that the success rate of finding the correct damage localization is increased when using multiple mode sets instead of a single mode set [15], [52], [27].

7.6.5. Transfer matrices-based statistical damage localization and quantification

Participants: Md Delwar Hossain Bhuyan, Michael Doehler, Laurent Mevel, Guillaume Gautier.

This work is in collaboration with GEM, Nantes and C. Ventura at UBC, Nantes.

Vibration measurements and a finite element model are used to locate loss of stiffness in a steel frame structure at the University of British Columbia. The Stochastic Dynamic Damage Locating Vector (SDDLV) is compared to a sensitivity based approach developed by the authors. Both approaches have in common to be built on the estimated transfer matrix difference between reference and damaged states. Both methods are tested for localization and quantification on a structure at University of British Columbia [26], [28].

7.6.6. Statistical damage localization based on Mahalanobis distance

Participant: Michael Doehler.

This work is in collaboration with Aalborg University, Structural Vibration Solutions and Universal Foundation in Denmark during the thesis of S. Gres (Aalborg University).

In this paper, a new Mahalanobis distance-based damage detection method is studied and compared to the wellknown subspace-based damage detection algorithm. Methods are implemented using control charts to enhance the resolution of the damage detection. The damage indicators are evaluated based on the ambient vibration signals from numerical simulations on a novel offshore support structure and experimental example of a full scale bridge. The results reveal that the performance of the two damage detection methods is similar, hereby implying merit of the new Mahalanobis distance-based approach, as it is less computationaly complex [32].

7.6.7. On the value of Information for SHM

Participant: Michael Doehler.

This work is issued from the COST Action TU1402.

The concept of value of information (VoI) enables quantification of the benefits provided by structural health monitoring (SHM) systems in principle. Its implementation is challenging, as it requires an explicit modelling of the structural system's life cycle, in particular of the decisions that are taken based on the SHM information. In this paper, we approach the VoI analysis through an influence diagram (ID), which supports the modelling process. We provide a simple example for illustration and discuss challenges associated with real-life implementation [39].

7.6.8. Structural system reliability and damage detection information

Participant: Michael Doehler.

This work is in collaboration with S. Thöns (DTU) during the thesis of L. Long (BAM).

This paper addresses the quantification of the value of damage detection system and algorithm information on the basis of Value of Information (VoI) analysis to enhance the benefit of damage detection information by providing the basis for its optimization before it is performed and implemented. The approach of the quantification the value of damage detection information builds upon the Bayesian decision theory facilitating the utilization of damage detection performance models, which describe the information and its precision on structural system level, facilitating actions to ensure the structural integrity and facilitating to describe the structural system performance and its functionality throughout the service life. The structural system performance is described with its functionality, its deterioration and its behavior under extreme loading. The structural system reliability given the damage detection information is determined utilizing Bayesian updating. The damage detection performance is described with the probability of indication for different component and system damage states taking into account type 1 and type 2 errors. The value of damage detection information is then calculated as the difference between the expected benefits and risks utilizing the damage detection information or not. With an application example of the developed approach based on a deteriorating Pratt truss system, the value of damage detection information is determined, demonstrating the potential of risk reduction and expected cost reduction [36].

7.6.9. Estimation of a cable resistance profile with readaptation of mismatched measurement instrument

Participants: Nassif Berrabah, Qinghua Zhang.

As the cumulative length of electric cables in modern systems is growing and as these systems age, it becomes of crucial importance to develop efficient tools to monitor the condition of wired connections. Therein, in contrast to hard faults (open or short circuits), the diagnosis of soft-faults requires a particular effort. Indeed, these faults are more difficult to detect, yet they are sometimes early warning signs of more important failures. In a previous paper, we proposed a method to compute the resistance profile of a cable from reflectometry measurements made at both ends of the cable. It enables detection, localization and estimation of dissipative soft-faults. In this reported work, we address the problem of impedance mismatch between the measurement instrument and the cable, based on a pre-processing of the measured data before running the estimation computations. It aims at reducing the impedance mismatch between instrumentation and the cable under test without physical intervention on the test fixtures. In addition, a measurement procedure has been developed

in order to get the two-ends reflectometry measurements without actually connecting both ends of the cable under test to a single instrument [25].

7.7. Sensor and hardware based research

7.7.1. Cracks detection in pavement by a distributed fiber optic sensing technology

Participant: Xavier Chapeleau.

This paper presents the feasibility of damage detection in asphalt pavements by embedded fiber optics as a new non-destructive inspection technique. The distributed fiber optic sensing technology based on the Rayleigh scattering was used in this study. The main advantage of this technique is that it allows to measure strains over a long length of fiber optic with a high spatial resolution, less than 1 cm. By comparing strain profiles measured at different times, an attempt was made to link strain changes with the appearance of damage (cracking) in the pavement. This non-destructive method was evaluated on accelerated pavement testing facility, in a bituminous pavement. In our experimentation, the optical fibers were placed near the bottom of the asphalt layer. The application of 728 000 heavy vehicle loads (65 kN dual wheel loads) was simulated in the experiment. Optical fiber measurements were made at regular intervals and surface cracking of the pavement was surveyed. After some traffic, a significant increase of strains was detected by the optical fibers at different points in the pavement structure, before any damage was visible. Later, cracking developed in the zones where the strain profiles were modified, thus indicating a clear relationship between the increased strains and crack initiation. These first tests demonstrate that distributed fiber optic sensors based on Rayleigh scattering can be used to detect crack initiation and propagation in pavements, by monitoring strain profiles in the bituminous layers [17].

7.7.2. Wireless sensors and GPS synchronization

Participants: Vincent Le Cam, David Pallier.

Most of recent development in WSN domain focused on energy (saving or harvesting), on wireless protocols, on embedded algorithms. But it is a fact that, most of monitoring applications need samples to be time-stamped. According to the application, the wished time resolution could be up to one second for automation monitoring, one millisecond for vibration, one microsecond for acoustic monitoring, one nanosecond for electricity or light propagation... The consequence for a Wireless network of electronic nodes is that, by nature, no common signal could physically provide a synchronization top. But, as each electronic device, a wireless sensor time-base uses a timer incremented by a quartz whose initial value is theoretical up to some p.p.m. and whose period drift on time because of age, temperature,... Two kind of solutions could be regarded : a synchronization signal provided by the wireless protocol itself; an absolute synchronization from a referential source such as: GPS, Frankfurt clock, Galileo,... In the first way, it will be demonstrated the poor accuracy and the need of energy such a mechanism offers. In the second way, the article will details how a deterministic (Universal Time), accurate and resilient algorithm has been implemented. The article also provides specific results of application on acoustic monitoring system and electricity propagation where the accuracy of a WSN has reached up to 10 nanosecond UT. Consequence on energy consumption of this algorithm are given with a description of future works to improve the energy balance while keeping the device sober and synchronized [33].

IPSO Project-Team

4. New Results

4.1. Multiscale numerical methods

4.1.1. Asymptotic preserving and time diminishing schemes for rarefied gas dynamic

In [10], we introduce a new class of numerical schemes for rarefied gas dynamic problems described by collisional kinetic equations. The idea consists in reformulating the problem using a micro-macro decomposition and successively in solving the microscopic part by using asymptotic preserving Monte Carlo methods. We consider two types of decompositions, the first leading to the Euler system of gas dynamics while the second to the Navier-Stokes equations for the macroscopic part. In addition, the particle method which solves the microscopic part is designed in such a way that the global scheme becomes computationally less expensive as the solution approaches the equilibrium state as opposite to standard methods for kinetic equations which computational cost increases with the number of interactions. At the same time, the statistical error due to the particle part of the solution decreases as the system approach the equilibrium state. This causes the method to degenerate to the sole solution of the macroscopic hydrodynamic equations (Euler or Navier-Stokes) in the limit of infinite number of collisions. In a last part, we will show the behaviors of this new approach in comparisons to standard Monte Carlo techniques for solving the kinetic equation by testing it on different problems which typically arise in rarefied gas dynamic simulations.

4.1.2. An exponential integrator for the drift-kinetic model

In [30], we propose an exponential integrator for the drift-kinetic equations in polar geometry. This approach removes the CFL condition from the linear part of the system (which is often the most stringent requirement in practice) and treats the remainder explicitly using Arakawa's finite difference scheme. The present approach is mass conservative, up to machine precision, and significantly reduces the computational effort per time step. In addition, we demonstrate the efficiency of our method by performing numerical simulations in the context of the ion temperature gradient instability. In particular, we find that our numerical method can take time steps comparable to what has been reported in the literature for the (predominantly used) splitting approach. In addition, the proposed numerical method has significant advantages with respect to conservation of energy and efficient higher order methods can be obtained easily. We demonstrate this by investigating the performance of a fourth order implementation.

4.1.3. Multiscale Particle-in-Cell methods and comparisons for the long-time two-dimensional Vlasov-Poisson equation with strong magnetic field

In [11], we applied different kinds of multiscale methods to numerically study the long-time Vlasov-Poisson equation with a strong magnetic field. The multiscale methods include an asymptotic preserving Runge-Kutta scheme, an exponential time differencing scheme, stroboscopic averaging method and a uniformly accurate two-scale formulation. We briefly review these methods and then adapt them to solve the Vlasov-Poisson equation under a Particle-in-Cell discretization. Extensive numerical experiments are conducted to investigate and compare the accuracy, efficiency, and long-time behavior of all the methods. The methods with the best performance under different parameter regimes are identified.

4.1.4. Nonlinear Geometric Optics based multiscale stochastic Galerkin methods for highly oscillatory transport equations with random inputs

In [31], we develop generalized polynomial chaos (gPC) based stochastic Galerkin (SG) methods for a class of highly oscillatory transport equations that arise in semiclassical modeling of non-adiabatic quantum dynamics. These models contain uncertainties, particularly in coefficients that correspond to the potentials of the molecular system. We first focus on a highly oscillatory scalar model with random uncertainty.

Our method is built upon the nonlinear geometrical optics (NGO) based method, developed in [12] for numerical approximations of deterministic equations, which can obtain accurate pointwise solution even without numerically resolving spatially and temporally the oscillations. With the random uncertainty, we show that such a method has oscillatory higher order derivatives in the random space, thus requires a frequency dependent discretization in the random space. We modify this method by introducing a new " time " variable based on the phase, which is shown to be non-oscillatory in the random space, based on which we develop a gPC-SG method that can capture oscillations with the frequency-independent time step, mesh size as well as the degree of polynomial chaos. A similar approach is then extended to a semiclassical surface hopping model system with a similar numerical conclusion. Various numerical examples attest that these methods indeed capture accurately the solution statistics pointwisely even though none of the numerical parameters resolve the high frequencies of the solution.

4.1.5. Nonlinear Geometric Optics method based multi-scale numerical schemes for highly-oscillatory transport equations

In [12], we introduce a new numerical strategy to solve a class of oscillatory transport PDE models which is able to capture accurately the solutions without numerically resolving the high frequency oscillations *in both space and time*. Such PDE models arise in semiclassical modeling of quantum dynamics with band-crossings, and other highly oscillatory waves. Our first main idea is to use the nonlinear geometric optics ansatz, which builds the oscillatory phase into an independent variable. We then choose suitable initial data, based on the Chapman-Enskog expansion, for the new model. For a scalar model, we prove that so constructed model will have certain smoothness, and consequently, for a first order approximation scheme we prove uniform error estimates independent of the (possibly small) wave length. The method is extended to systems arising from a semiclassical model for surface hopping, a non-adiabatic quantum dynamic phenomenon. Numerous numerical examples demonstrate that the method has the desired properties.

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

In [5], 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.

4.1.7. A particle micro-macro decomposition based numerical scheme for collisional kinetic equations in the diffusion scaling

In [29], we derive particle schemes, based on micro-macro decomposition, for linear kinetic equations in the diffusion limit. Due to the particle approximation of the micro part, a splitting between the transport and the collision part has to be performed, and the stiffness of both these two parts prevent from uniform stability. To overcome this difficulty, the micro-macro system is reformulated into a continuous PDE whose coefficients are no longer stiff, and depend on the time step Δt in a consistent way. This non-stiff reformulation of the micro-macro system allows the use of standard particle approximations for the transport part, and extends a previous work where a particle approximation has been applied using a micro-macro decomposition on kinetic equations in the fluid scaling. Beyond the so-called asymptotic-preserving property which is satisfied by our schemes, they significantly reduce the inherent noise of traditional particle methods, and they have a computational cost which decreases as the system approaches the diffusion limit.

4.1.8. Uniformly accurate forward semi-Lagrangian methods for highly oscillatory Vlasov-Poisson equation

This work [13] is devoted to the numerical simulation of a Vlasov-Poisson equation modeling charged particles in a beam submitted to a highly oscillatory external electric field. A numerical scheme is constructed for this model. This scheme is uniformly accurate with respect to the size of the fast time oscillations of the solution, which means that no time step refinement is required to simulate the problem. The scheme combines the forward semi-Lagrangian method with a class of Uniformly Accurate (UA) time integrators to solve the characteristics. These UA time integrators are derived by means of a two-scale formulation of the characteristics, with the introduction of an additional periodic variable. Numerical experiments are done to show the efficiency of the proposed methods compared to conventional approaches.

4.1.9. Uniformly accurate multiscale time integrators for second order oscillatory differential equations with large initial data

In [23], we apply the modulated Fourier expansion to a class of second order differential equations which consists of an oscillatory linear part and a nonoscillatory nonlinear part, with the total energy of the system possibly unbounded when the oscillation frequency grows. We comment on the difference between this model problem and the classical energy bounded oscillatory equations. Based on the expansion, we propose the multiscale time integrators to solve the ODEs under two cases: the nonlinearity is a polynomial or the frequencies in the linear part are integer multiples of a single generic frequency. The proposed schemes are explicit and efficient. The schemes have been shown from both theoretical and numerical sides to converge with a uniform second order rate for all frequencies. Comparisons with popular exponential integrators in the literature are done.

4.1.10. Unconditional and optimal H^2-error estimates of two linear and conservative finite difference schemes for the Klein-Gordon-Schrödinger equation in high dimensions

In [21], we focus on the optimal error bounds of two finite difference schemes for solving the d-dimensional ($d = 2, 3$) nonlinear Klein-Gordon-Schrödinger (KGS) equations. The proposed finite difference schemes not only conserve the mass and energy in the discrete level but also are efficient in practical computation because only two linear systems need to be solved at each time step. Besides the standard energy method, an induction argument as well as a 'lifting' technique are introduced to establish rigorously the optimal H^2-error estimates without any restrictions on the grid ratios, while the previous works either are not rigorous enough or often require certain restriction on the grid ratios. The convergence rates of the proposed schemes are proved to beat $O(h^2 + \tau^2)$ with mesh size h and time step τ in the discrete H^2-norm. The analysis method can be directly extended to other linear finite difference schemes for solving the KGS equations in high dimensions. Numerical results are reported to confirm the theoretical analysis for the proposed finite difference schemes

4.1.11. A combination of multiscale time integrator and two-scale formulation for the nonlinear Schrödinger equation with wave operator

In [22], we consider the nonlinear Schrödinger equation with wave operator (NLSW), which contains a dimensionless parameter $0 < \varepsilon \leq 1$. As $0 < \varepsilon << 1$, the solution of the NLSW propagates fast waves in time with wavelength $O(\varepsilon^2)$ and the problem becomes highly oscillatory in time. The oscillations come from two parts. One part is from the equation and another part is from the initial data. For the ill-prepared initial data case as described in Bao and Cai (2014) which brings inconsistency in the limit regime, standard numerical methods have strong convergence order reduction in time when becomes small. We review two existing methods to solve the NLSW: an exponential integrator and a two-scale method. We comment on their order reduction issues. Then we derive a multiscale decomposition two-scale method for solving the NLSW by first performing a multiscale decomposition on the NLSW which decomposes it into a well-behaved part and an energy-unbounded part, and then applying an exponential integrator for the well-behaved part and a two-scale approach for the energy-unbounded part. Numerical experiments are conducted to test the proposed method which shows uniform second order accuracy without significant order reduction for all $0 < \varepsilon \leq 1$. Comparisons are made with the existing methods.

4.1.12. Uniformly accurate numerical schemes for the nonlinear Dirac equation in the nonrelativistic limit regime

In [18], we apply the two-scale formulation approach to propose uniformly accurate (UA) schemes for solving the nonlinear Dirac equation in the nonrelativistic limit regime. The nonlinear Dirac equation involves two small scales ε and ε^2 with epsilon $\to 0$ in the nonrelativistic limit regime. The small parameter causes high oscillations in time which brings severe numerical burden for classical numerical methods. We transform our original problem as a two-scale formulation and present a general strategy to tackle a class of highly oscillatory problems involving the two small scales ε and ε^2. Suitable initial data for the two-scale formulation is derived to bound the time derivatives of the augmented solution. Numerical schemes with uniform (with respect to $\varepsilon \in (0; 1]$) spectral accuracy in space and uniform first order or second order accuracy in time are proposed. Numerical experiments are done to confirm the UA property.

4.1.13. A formal series approach to the center manifold theorem

In [6], we consider near-equilibrium systems of ordinary differential equations with explicit separation of the slow and stable manifolds. Formal B-series like those previously used to analyze highly-oscillatory systems or to construct modified equations are employed here to construct expansions of the change of variables, the center invariant manifold and the reduced model. The new approach may be seen as a process of reduction to a normal form, with the main advantage, as compared to the standard view conveyed by the celebrated center manifold theorem, that it is possible to recover the complete solution at any time through an explicit change of variables.

4.1.14. Convergence of multi-revolution composition time-splitting methods for highly oscillatory differential equations of Schrödinger type

In [8], the convergence behaviour of multi-revolution composition methods combined with time-splitting methods is analysed for highly oscillatory linear differential equations of Schrödinger type. Numerical experiments illustrate and complement the theoretical investigations.

4.1.15. Highly-oscillatory evolution equations with multiple frequencies: averaging and numerics

In [7], we are concerned with the application of the recently introduced multi-revolution composition methods, on the one hand, and two-scale methods, on the other hand, to a class of highly-oscillatory evolution equations with multiple frequencies. The main idea relies on a well-balanced reformulation of the problem as an equivalent mono-frequency equation which allows for the use of the two aforementioned techniques.

4.1.16. Optimality and resonances in a class of compact finite difference schemes of high order

In [25], we revisit the old problem of compact finite difference approximations of the homogeneous Dirichlet problem in dimension 1. We design a large and natural set of schemes of arbitrary high order, and we equip this set with an algebraic structure. We give some general criteria of convergence and we apply them to obtain two new results. On the one hand, we use Padé approximant theory to construct, for each given order of consistency, the most efficient schemes and we prove their convergence. On the other hand, we use diophantine approximation theory to prove that almost all of these schemes are convergent at the same rate as the consistency order, up to some logarithmic correction.

4.2. mathematical analysis of multiscale partial differential equations

4.2.1. Collision of almost parallel vortex filaments

In [3], we investigate the occurrence of collisions in the evolution of vortex filaments through a system introduced by Klein, Majda and Damodaran and 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 a previous

work. In the second part of this paper we extend this construction to the case of an arbitrary number of filaments, with polygonial symmetry, that are perturbations of a configuration of parallel vortex filaments forming a polygon, with or without its center, rotating with constant angular velocity.

4.2.2. Free vibrations of axisymmetric shells: parabolic and elliptic cases

In [9], approximate eigenpairs (quasimodes) of axisymmetric thin elastic domains with laterally clamped boundary conditions (Lamé system) are determined by an asymptotic analysis as the thickness (2ε) tends to zero. The departing point is the Koiter shell model that we reduce by asymptotic analysis to a scalar model that depends on two parameters: the angular frequency k and the half-thickness ε. Optimizing k for each chosen ε, we find power laws for k in function of ε that provide the smallest eigenvalues of the scalar reductions. Corresponding eigenpairs generate quasimodes for the 3D Lamé system by means of several reconstruction operators, including boundary layer terms. Numerical experiments demonstrate that in many cases the constructed eigenpair corresponds to the first eigenpair of the Lamé system. Geometrical conditions are necessary to this approach: The Gaussian curvature has to be nonnegative and the azimuthal curvature has to dominate the meridian curvature in any point of the midsurface. In this case, the first eigenvector admits progressively larger oscillation in the angular variable as ε tends to 0 . Its angular frequency exhibits a power law relation of the form $k = \gamma\varepsilon^{\beta}$ with $\beta = \frac{1}{4}$ in the parabolic case (cylinders and trimmed cones), and the various β's ($\frac{2}{5}$, $\frac{3}{7}$ and $\frac{1}{3}$ in the elliptic case. For these cases where the mathematical analysis is applicable, numerical examples that illustrate the theoretical results are presented.

4.2.3. High frequency oscillations of first eigenmodes in axisymmetric shells as the thickness tends to zero

In [24], the lowest eigenmode of thin axisymmetric shells is investigated for two physical models (acoustics and elasticity) as the shell thickness (2ε) tends to zero. Using a novel asymptotic expansion we determine the behavior of the eigenvalue $\lambda(\varepsilon)$ and the eigenvector angular frequency $k(\varepsilon)$ for shells with Dirichlet boundary conditions along the lateral boundary, and natural boundary conditions on the other parts. First, the scalar Laplace operator for acoustics is addressed, for which $k(\varepsilon)$ is always zero. In contrast to it, for the Lamé system of linear elasticity several different types of shells are defined, characterized by their geometry, for which $k(\varepsilon)$ tends to infinity as ε tends to zero. For two families of shells: cylinders and elliptical barrels we explicitly provide $\lambda(\varepsilon)$ and $k(\varepsilon)$ and demonstrate by numerical examples the different behavior as ε tends to zero.

4.2.4. Semiclassical Sobolev constants for the electro-magnetic Robin Laplacian

This paper [15] is devoted to the asymptotic analysis of the optimal Sobolev constants in the semiclassical limit and in any dimension. We combine semiclassical arguments and concentration-compactness estimates to tackle the case when an electromagnetic field is added as well as a smooth boundary carrying a Robin condition. As a byproduct of the semiclassical strategy, we also get exponentially weighted localization estimates of the minimizers.

4.2.5. On the MIT Bag Model in the Non-relativistic Limit

This paper [2] is devoted to the spectral investigation of the MIT bag model, that is, the Dirac operator on a smooth and bounded domain of $\mathbb{R}^3$ with certain boundary conditions. When the mass m goes to $\pm\infty$, we provide spectral asymptotic results.

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

In [4], 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.

4.2.7. Nonlinear stability criteria for the HMF Model

In [17], we study the nonlinear stability of a large class of inhomogeneous steady state solutions to the Hamiltonian Mean Field (HMF) model. Under a specific 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.

4.2.8. Strong confinement limit for the nonlinear Schrödinger equation constrained on a curve

This paper [20] is devoted to the cubic nonlinear Schrödinger equation in a two dimensional waveguide with shrinking cross section. For a Cauchy data living essentially on the first mode of the transverse Laplacian, we provide a tensorial approximation of the solution in this limit, with an estimate of the approximation error, and derive a limiting nonlinear Schrödinger equation in dimension one.

4.2.9. Stable ground states for the HMF Poisson Model

In [36], we prove the nonlinear orbital stability of a large class of steady states solutions to the Hamiltonian Mean Field (HMF) system with a Poisson interaction potential. These steady states are obtained as minimizers of an energy functional under one, two or infinitely many constraints. The singularity of the Poisson potential prevents from a direct run of the general strategy which was based on generalized rearrangement techniques, and which has been recently extended to the case of the usual (smooth) cosine potential. Our strategy is rather based on variational techniques. However, due to the boundedness of the space domain, our variational problems do not enjoy the usual scaling invariances which are, in general, very important in the analysis of variational problems. To replace these scaling arguments, we introduce new transformations which, although specific to our context, remain somehow in the same spirit of rearrangements tools introduced in the references above. In particular, these transformations allow for the incorporation of an arbitrary number of constraints, and yield a stability result for a large class of steady states.

4.2.10. The quantum Liouville-BGK equation and the moment problem

This work [19] is devoted to the analysis of the quantum Liouville-BGK equation. This equation arises in the work of Degond and Ringhofer on the derivation of quantum hydrodynamical models from first principles. Their theory consists in transposing to the quantum setting the closure strategy by entropy minimization used for kinetic equations. The starting point is the quantum Liouville-BGK equation, where the collision term is defined via a so-called quantum local equilibrium, defined as a minimizer of the quantum free energy under a local density constraint. We then address three related problems: we prove new results about the regularity of these quantum equilibria; we prove that the quantum Liouville-BGK equation admits a classical solution; and we investigate the long-time behavior of the solutions. The core of the proofs is based on a fine analysis of the properties of the minimizers of the free energy.

4.2.11. Averaging of nonlinear Schrödinger equations with strong magnetic confinement

In [16], we consider the dynamics of nonlinear Schrödinger equations with strong constant magnetic fields. In an asymptotic scaling limit the system exhibits a purely magnetic confinement, based on the spectral properties of the Landau Hamiltonian. Using an averaging technique we derive an associated effective description via an averaged model of nonlinear Schrödinger type. In a special case this also yields a derivation of the LLL equation.

4.3. mathematical analysis of stochastic partial differential equations

4.3.1. Large deviations for the dynamic Φ_d^{2n} model

In [27], we are dealing with the validity of a large deviation principle for a class of reaction-diffusion equations with polynomial non-linearity, perturbed by a Gaussian random forcing. We are here interested in the regime

where both the strength of the noise and its correlation are vanishing, on a length scale ρ and $\delta(\rho)$, respectively, with $0 < \rho, \delta(\rho) << 1$. We prove that, under the assumption that ρ and $\delta(\rho)$ satisfy a suitable scaling limit, a large deviation principle holds in the space of continuous trajectories with values both in the space of square-integrable functions and in Sobolev spaces of negative exponent. Our result is valid, without any restriction on the degree of the polynomial nor on the space dimension.

4.3.2. Solution to the stochastic Schrödinger equation on the full space

In [33], we show how the methods recently applied by Debussche and Weber to solve the stochastic nonlinear Schrödinger equation on $\mathbb{T}^2$ can be enhanced to yield solutions on $\mathbb{R}^2$ if the non-linearity is weak enough. We prove that the solutions remains localized on compact time intervals which allows us to apply energy methods on the full space.

4.3.3. A law of large numbers in the supremum norm for a multiscale stochastic spatial gene network

In [34], we study the asymptotic behavior of multiscale stochastic spatial gene networks. Multiscaling takes into account the difference of abundance between molecules, and captures the dynamic of rare species at a mesoscopic level. We introduce an assumption of spatial correlations for reactions involving rare species and a new law of large numbers is obtained. According to the scales, the whole system splits into two parts with different but coupled dynamics. The high scale component converges to the usual spatial model which is the solution of a partial differential equation, whereas, the low scale component converges to the usual homogeneous model which is the solution of an ordinary differential equation. Comparisons are made in the supremum norm.

4.3.4. Long time behavior of Gross-Pitaevskii equation at positive temperature

In [32], the stochastic Gross-Pitaevskii equation is used as a model to describe Bose-Einstein condensation at positive temperature. The equation is a complex Ginzburg Landau equation with a trapping potential and an additive space-time white noise. Two important questions for this system are the global existence of solutions in the support of the Gibbs measure, and the convergence of those solutions to the equilibrium for large time. In this paper, we give a proof of these two results in one space dimension. In order to prove the convergence to equilibrium, we use the associated purely dissipative equation as an auxiliary equation, for which the convergence may be obtained using standard techniques.

4.3.5. An integral inequality for the invariant measure of a stochastic reaction–diffusion equation

In [14], 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 $\mathcal{H}$.

4.3.6. Kolmogorov equations and weak order analysis for SPDES with nonlinear diffusion coefficient

In [26], we provide new regularity results for the solutions of the Kolmogorov equation associated to a SPDE with nonlinear diffusion coefficients and a Burgers type nonlinearity. This generalizes previous results in the simpler cases of additive or affine noise. The basic tool is a discrete version of a two sided stochastic integral which allows a new formulation for the derivatives of these solutions. We show that this can be used to generalize the weak order analysis performed by Debussche in 2011. The tools we develop are very general and can be used to study many other examples of applications.

4.3.7. Approximation-diffusion in stochastically forced kinetic equations

In [35], we derive the hydrodynamic limit of a kinetic equation where the interactions in velocity are modelled by a linear operator (Fokker-Planck or Linear Boltzmann) and the force in the Vlasov term is a stochastic process with high amplitude and short-range correlation. In the scales and the regime we consider, the hydrodynamic equation is a scalar second-order stochastic partial differential equation. Compared to the deterministic case, we also observe a phenomenon of enhanced diffusion.

DYLISS Project-Team

7. New Results

7.1. Data integration and pre-processing with semantic-based technologies

Participants: Olivier Dameron, Xavier Garnier, Yann Rivault, Anne Siegel, Denis Tagu.

Interoperable infrastructure and implementation of a health data model for remote monitoring of chronic diseases with comorbidities In the context of **telemedecine**, we worked on a numerical application for monitoring patients with chronic diseases. We have developed a system based on a formal ontology that integrates the alert information and the patient data extracted from the electronic health record in order to better classify the importance of alerts. A pilot study was conducted on atrial fibrillation alerts. The results suggest that this approach has the potential to significantly reduce the alert burden in telecardiology [101], [100]. In 2017, we proposed an architecture supporting data exchange in the context of multiple chronic diseases [*O. Dameron, Y. Rivault*] [27].

AskOmics, a web tool to integrate and query biological data using semantic web technologies The software AksOmics has been adapted to two types of scientific topics important in agronomical and environmental sciences: plant genomic data and insect pest genomic data. With *AskOmics*, plant genomicists (from academic and private labs from the Rapsodyn project - Investment for the future) working on the rapeseed (Brassica napus) are able to tackle the understanding of which gene copy is active or repressed in key developmental processes in relation with seed quality and oil production, in the frame of plant breeding. Additionally, entomologists use this tool to extract valuable knowledge on the way insect pests such as aphids are able to rapidly disseminate on crops, in the frame of free-pesticide methods for plant protection. *AskOmics* has been presented to the international community of insect genomics (i5k: http://i5k.github.io/) by web-seminars and *AskOmics* developers have been invited at international workshops. For facilitating AskOmics's adoption by end-user, it has recently been integrated within the Galaxy workflow engine [*O. Dameron, X. Garnier, A. Siegel, D. Tagu*] [28], [29], [23]

7.2. Data and knowledge integration based on combinatorial optimization

Participants: Meziane Aite, Lucas Bourneuf, Marie Chevallier, Damien Eveillard, Clémence Frioux, Jeanne Got, Julie Laniau, François Moreews, Jacques Nicolas, Anne Siegel.

A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs Our work on the identification of upstream regulators within large-scale knowledge databases (prototype *KeyRegulatorFinder*) [59] was valuable for figuring out the main gene-regulators of the response of porks to several diets [*F. Moreews, A. Siegel*] [18]

FCA in a Logical Programming Setting for Visualization-oriented Graph Compression We have explored the underlying idea of lossless **network compression** to address the problem of uncertainty in biological networks built from predictions, to help to visualize the networks and to classify their nodes in accordance with available annotations [119]. Network compression has been used with success in Dresden (M. Schroeder) with a heuristic approach called Power Graph analysis building abstract graphs where nodes are clusters of nodes in the initial graph and edges represent bicliques between two sets of nodes. First encouraging results have been presented (best paper award) showing that it is possible to mimic the Power Graph behaviour while opening the possibility to achieve better compression levels compared to alternative compression schema. [*L. Bourneuf, J. Nicolas*] [24]

Metabolic network completion and analysis We released the application paper of the tool *Meneco*, a tool dedicated to the topological gap-filling of genome-scale draft metabolic networks. The tool reformulates gap-filling as a qualitative combinatorial optimization problem, omitting constraints raised by the stoichiometry, and solves this problem using Answer Set Programming. Run on an artificial test set of 10,800 degraded *Escherichia coli* networks, we evidenced that *Meneco* outperforms the stoichiometry-based tool *Gapfill* in terms of precision. In addition, *Meneco* reports 10 times less putative reactions than MILP-based tool *Fastgapfill* for an equivalent precision. This is a strong advantage for manual curation post-processing, since curating 50 to 80 reactions is still possible whereas manually-curating 800 reactions is out-of-range. Meneco was applied to the reconstruction and understanding of a pathogeneic strain of salmon. [*C. Frioux, J. Got, A. Siegel*] [21], [16]

Toward the study of metabolic functions in communities of organisms In [21], we provided a first example on how to use topological metabolic modeling to assess the complementarity between two members of an algal ecosystem. Since this study, we generalized the selection of subcommunities of interest and propose likely interactions that could occur between seaweeds and their associated bacteria. A focus has also been done on plant microbiota and the reasons underlying the organization of the community. Altogether, these on-going works enable a better understanding of holobiont organizations and functioning. [*M. Aite, M. Chevallier, C. Frioux, J. got, A. Siegel, C. Trottier*] [21], [31], [30]

Hybrid Metabolic Network Completion In order to improve the precision of gap-filling approaches, we introduced a hybrid approach to formally reconcile existing stoichiometric and topological approaches to network completion in a unified formalism. An hybrid ASP encoding based on MILP constraint propagator was developed. It relies upon the theory reasoning capacities of the ASP system Clingo to solve the resulting logic program with linear constraints over reals. For short, this technology made it possible to combine the best of the combinatorial problem solver Clingo with the MILP solver CPlex. Run on the artificial test set of 10,800 degraded *Escherichia coli* networks introduced in [21], our approach yielded greatly superior results than obtainable from purely qualitative or MILP approaches. [*C. Frioux, A. Siegel*] [26], [19]

Combining graph and flux-based structures to decipher phenotypic essential metabolites within metabolic networks Whenever flux or graph-based criteria are used to study metabolic networks, these analyses are generally centered on the outcome of the network and considers all metabolic compounds to be equivalent in this respect. We generalized the concept of essentiality to metabolites and introduced the concept of the phenotypic essential metabolite (PEM) which influences the growth phenotype according to sustainability, producibility or optimal-efficiency criteria. The exhaustive study of phenotypic essential metabolites in six genome-scale metabolic models suggests that the combination and the comparison of graph, stoichiometry and optimal flux-based criteria allow some features of the metabolic network functionality to be deciphered by focusing on a small number of compounds. [*C. Frioux, J. Laniau, A. Siegel*] [19]

7.3. Systems biology

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

A modeling approach to evaluate the balance between bioactivation and detoxification of MeIQx in human hepatocytes Heterocyclic aromatic amines (HAA), including MeIQx, are environmental and food contaminants that are potentially carcinogenic for humans. Using a computational approach, we developed a numerical model for MeIQx metabolism that predicts the MeIQx biotransformation into detoxification or bioactivation pathways according to the concentration of MeIQx. Our results demonstrate that CYP1A2 is a key enzyme in the system that regulates the balance between bioactivation and detoxification. This highlights the importance of complex regulations of enzyme competitions that should be taken into account in any multi-organ model [*V. Delannée, A. Siegel, N. Théret*] [17]

caspo: a toolbox for automated reasoning on the response of logical signaling networks families The accompanying paper of the complete family of modules introduced in the caspo software was published in 2017 (see software section for details) [*A. Siegel*] [22]

Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering: Application to TGF-β Signaling At a dynamical level, in [40], reaction-based and regulatory information was transpoed in a unified formalism of enriched Petri Nets (discrete dynamical systems), namely a simplified version of guarded transitions in which we introduced temporal parameters for each transition to manage competition and cooperation between parts of the models. This allowed integrating the 137 human signaling maps from the Pathway Interaction Database (PID) into a single unified large-scale dynamic model. Simulation and model checking analyses evidence that 15,934 different sets of molecules are able to regulate 159 of TGF-β target genes (TGF-β is a multifunctional cytokine that regulates mammalian cell development, differentiation, and homeostasis). Further analysis of these 15,934 sets of molecules by biological experts is obviously impractical. Our study identified five clusters of sets of molecules for which enrichment analysis highlighted the over-represented molecules as well as the specific biological processes they are associated with. These results are biologically-relevant and consistent with the pleiotropic nature of TGF-β [*J. Coquet, N. Théret, O. Dameron*] [25]

A Logic for checking the probabilistic steady-state properties of reaction networks. We have constructed a probabilistic analog to flux balance analysis of reaction networks to enable a formal verification of logical constrains about the stationary regime of a system by using information from experimental variances and co-variances. This is mainly based on a stationary analysis of the probabilistic dynamics relying on a Bernoulli approximation of a reaction network. The analysis requires solving non linear optimization problems [*J. Bourdon, A. Siegel*] [20]

7.4. Sequence and structure annotation

Participants: Catherine Belleannée, François Coste, Jacques Nicolas.

Better scoring schemes for the recognition of functional proteins by protomata The machine learning algorithm included in *Protomata-learner* learns weighted automata representing both functional families from the sequences of amino acids, and the possible disjunctions between members. We investigated alternative sequence weighting strategies and null-models. We introduced a normalization of the score, and a method to assess the significance of scores, to simplify the prediction. Preliminary results show a good improvement of the prediction power of the computed models. [*F. Coste*] [36]

Detection of mutated primers and impact on targeted metagenomics results In targeted metagenomics, an initial task is the detection in each sequence of the primers used for amplifying the targeted region. The selected sequences are then trimmed and clustered in order to inventory the species present in the sample. Common pratices consist in retaining only the sequences with perfect primers (i.e. non-mutated by sequencing error). In the context of a study characterizing the biodiversity of tropical soils in unicellular eukaryotes, we have implemented the search for mutated primers, using the grammatical pattern matching tool Logol, and shown that retrieving sequences with mutated primers has a significant impact on targeted metagenomics results, as it makes possible to detect more species (7% additional OTUs in our study). [*C. Belleannée*] [34].

First landscape of binding to chromosomes for a domesticated mariner transposase in the human genome. In order to study the diversity of genomic targets of the SETMAR protein in two colorectal cell lines, a first task was to massively discover the Made1 80-bp transposon element in the human genome. For that, we used our Logol grammar-like approach to look for non perfect Made1 instances. In Logol, a pattern can be divided into several sub-patterns. The Made1 model took advantage of this feature to strengthen the most conserved regions. Cumulating this search with the Blast alignment search permitted to significantly increase the Made1 annotation in the human genome.[*C. Belleannée*] [33]

FLUMINANCE Project-Team

7. New Results

7.1. Fluid motion estimation

7.1.1. Stochastic uncertainty models for motion estimation

Participants: Shengze Cai, Etienne Mémin, Musaab Khalid Osman Mohammed.

The objective consists here in relying on a stochastic transport formulation to propose a luminance conservation assumption dedicated to the measurement of large-scale fluid flows velocity. This formulation, relying on the modeling under location uncertainty principle developed in the team, has the great advantage to incorporate from the beginning an uncertainty on the unresolved (turbulent) motion measurement. This uncertainty modeled as a possibly inhomogeneous random field uncorrelated in time 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. It provides also a reinterpretation, in terms of uncertainty, of classical regularization functionals proposed in the context of motion estimation. Nevertheless, at variance to classical motion estimation methods, this approach enables to estimate the so-called regularization parameter, which is in our framework related to the variance of the unresolved component of motion component. The resulting parameter-free estimator has shown to outperform state-of-the-art results of the literature [14]. This kind of method is applied on turbulent flows and in the context of river hydrologic applications through a collaboration with the Irstea Lyon research group (HHLY). A method for the 3D reconstruction of the river plane has been also proposed in this context.This study is performed within the PhD thesis of Musaab Mohammed.

7.1.2. Surface Currents estimation from Shore-Based Videos

Participant: Pierre Derian.

A wavelet based motion estimator has been extended for the recovery of instantaneous fields of surface current from shore-based and unmanned aerial vehicle videos. This study published in [16] and [34] demonstrated clearly the high potential of this method in the nearshore, where the rapid development of webcams and drones offers a large amount of applications for swimming and surfing safety, engineering and naval security and research purpose, by providing quantitative information. This work has been conducted within a collaboration with the Legos laboratory.

7.1.3. Development of an image-based measurement method for large-scale characterization of indoor airflows

Participants: Dominique Heitz, Etienne Mémin, Romain Schuster.

The goal is to design a new image-based flow measurement method for large-scale industrial applications. From this point of view, providing *in situ* measurement technique requires: (i) the development of precise models relating the large-scale flow observations to the velocity; (ii) appropriate large-scale regularization strategies; and (iii) adapted seeding and lighting systems, like Hellium Filled Soap Bubles (HFSB) and led ramp lighting. This work conducted within the PhD of Romain Schuster in collaboration with the compagny ITGA has started in february 2016. The first step has been to evaluate the performances of a stochastic uncertainty motion estimator when using large scale scalar images, like those obtained when seeding a flow with smoke. The PIV characterization of flows on large fields of view requires an adaptation of the motion estimation method from image sequences. The backward shift of the camera coupled to a dense scalar seeding involves a large scale observation of the flow, thereby producing uncertainty about the observed phenomena. By introducing a stochastic term related to this uncertainty into the observation term, we obtained a significant improvement of the estimated velocity field accuracy [41].

7.1.4. 3D flows reconstruction from image data

Participants: Dominique Heitz, Cédric Herzet.

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 continued our investigation on the problem of efficient volume reconstruction. During the last years, we have focussed our attention on several families of convex optimization algorithms allowing to accelerate the standard procedures encountered in the Tomo-PIV literature while accounting for the non-negativity and the sparsity of the sought solutions. So far, the assessment of the proposed algorithms were exclusively done on synthetic data. This year, we started the process of validating the proposed procedures on real experimental data.

We started through a collaboration with Irstea to study ensemble assimilation methods for the fast reconstruction of 3D tomo-PIV motion field. The approach relies on a simplified dynamics of the flow and is a generalization of one of the popular emerging flow reconstruction technique of the PIV community referred to as "Shake the box". The study will be developed within an Irstea post-doctoral funding.

7.1.5. Sparse-representation algorithms

Participant: Cédric Herzet.

The paradigm of sparse representations is a central concept 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. This year, we continued our investigations on “screening” methodologies, that is procedures allowing for the rapid identification of (some of) the zeros of the sought sparse vector. More specifically, we designed low-complexity procedures enabling to screen groups of atoms by only performing one single test. This work has been submitted to the IEEE international conference on acoustic, speech and signal processing (ICASSP).

7.2. Tracking, Data assimilation and model-data coupling

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

Participants: Mohamed Yacine Ben Ali, Pranav Chandramouli, Dominique Heitz, Etienne Mémin.

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 is experimented on shear layer flows and on wake flows generated on the wind tunnel of Irstea Rennes. Within this study we aim 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. A 4DVar assimilation technique based on the numerical code Incompact3D has been implemented for that purpose to control the inlet and initial conditions in order to reconstruct a turbulent wake flow behind an unknown obstacle. We are actually extending this first data assimilation technique to control the subgrid parameters. This study is performed in collaboration with Sylvain Laizet (Imperial College). In another axis of research, in collaboration with the CSTB Nantes centre and within the PhD of Yacine Ben Ali we will explore the definition of efficient data assimilation schemes for wind engineering. The goal will be here to couple Reynolds average model to pressure data at the surface of buildings. The final purpose will consist in proposing improved data-driven simulation models for architects.

7.2.2. Ensemble variational data assimilation of large-scale 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. Good results have been obtained for simulation under location uncertainty of 1D and 2D shallow water models [26]. This open the way to the definition of efficient data assimilation schemes for the coupling of high resolution data with large scale dynamical system.

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

Participants: Mamadou Diallo, Dominique Heitz, Cédric Herzet, Etienne Mémin, Valentin Resseguier.

During the PhD thesis of Valentin Resseguier 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 POD-Galerkin methods. This uncertainty modeling methodology provides a theoretically grounded technique to define an appropriate subgrid tensor as well drift correction terms. The pertinence of this stochastic reduced order model has been successfully assessed on several wake flows at different Reynold number. It has been shown to be much more stable than the usual reduced order model construction techniques. Beyond the definition of a stable reduced order model, the modeling under location uncertainty paradigm offers a unique way to analyse from the data of a turbulent flow the action of the small-scale velocity components on the large-scale flow [25]. Regions of prominent turbulent kinetic energy, direction of preferential diffusion, as well as the small-scale induced drift can be identified and analyzed to decipher key players involved in the flow. This study has been published in the journal of fluid mechanics. Note that these reduced order models can be extended to a full system of stochastic differential equation driving all the temporal modes of the reduced system (and not only the small-scale modes). This full stochastic system has been evaluated on wake flow at moderate Reynolds number. For this flow the system has shown to provide very good uncertainty quantification properties as well meaningful physical behavior with respect to the simulation of the neutral modes of the dynamics. This study described in the PhD of Valentin Resseguier will be soon submitted to a journal paper.

On the other hand, in the following of several approaches proposed by the team [49], [53], we continued our investigations on the estimation of deterministic reduced order model from partial observations. In this line of research, we proposed a Bayesian framework for the construction of reduced-order models from image data. Our framework combines observation and prior information on the system to reduce the model and takes into account the uncertainties on the parameters of the model. The proposed approach reduces to some well-known model-reduction techniques for complete observations (i.e., the observation operator can be inverted). A theoretical analysis of our methodology has been investigated 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 led to publications in a journal [18] and a conference [33].

7.3. Analysis and modeling of turbulent flows and geophysical flows

7.3.1. Geophysical flows modeling under location uncertainty

Participants: Pierre Derian, Long Li, Etienne Mémin, Valentin Resseguier.

In this research axis we have devised a principle to derive representation of flow dynamics under uncertainty. Such an uncertainty is formalized through the introduction of a random term that enables 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 partially known forcing. Rigorously derived from a stochastic version of the Reynolds transport theorem [9], this framework, referred to as modeling under location uncertainty, encompasses several meaningful mechanisms for turbulence modeling. It indeed introduces without any supplementary assumption the following pertinent mechanisms for turbulence modeling: (i) a dissipative operator related to the mixing effect of the large-scale components by the small-scale velocity; (ii) a multiplicative noise representing small-scale energy backscattering; and (iii) a modified advection term related to the so-called *turbophoresis* phenomena, attached to the migration of inertial particles in regions of lower turbulent diffusivity.

In a series of papers we have shown how such modeling can be applied to provide stochastic representations of a variety of classical geophysical flows dynamics [24], [23], [22]. Numerical simulations and uncertainty quantification have been performed on Quasi Geostophic approximation (QG) of oceanic models. It has been shown that such models lead to remarkable estimation of the unresolved errors at variance to classical eddy viscosity based models. The noise brings also an additional degree of freedom in the modeling step and pertinent diagnostic relations and variations of the model can be obtained with different scaling assumptions of the turbulent kinetic energy (i.e. of the noise amplitude). The performances of such systems have been assessed also on an original stochastic representation of the Lorenz 63 derived from the modeling under location uncertainty [15]. In this study it has been shown that the stochastic version enabled to explore in a much faster way the region of the deterministic attractor. This effort has been undertaken within a fruitful collaboration with Bertrand Chapron (LOPS/IFREMER). In the PhD of Long Li, starting this year, we will continue this effort. The goal will be to propose relevant techniques to define or calibrate the noise term from data. In that prospect, we intend to explore statistical learning techniques.

7.3.2. Large eddies simulation models under location uncertainty

Participants: Mohamed Yacine Ben Ali, Pranav Chandramouli, Dominique Heitz, Etienne Mémin.

The models under location uncertainty recently introduced by Mémin (2014) [9] provide a new outlook on LES modeling for turbulence studies. These models are derived from a stochastic transport principle. The associated stochastic conservation equations are similar to the filtered Navier- Stokes equation wherein we observe a sub-grid scale dissipation term. However, in the stochastic version, an extra term appears, termed as "velocity bias", which can be treated as a biasing/modification of the large-scale advection by the small scales. This velocity bias, introduced artificially in the literature, appears here automatically through a decorrelation assumption of the small scales at the resolved scale. All sub-grid contributions for the stochastic models are defined by the small-scale velocity auto-correlation tensor. This large scale modeling has been assed and compared to several classical large-scale models on several flows, namely a flow over a circular cylinder at Re $\sim$ 3900 [32], a smooth channel flow at Re(tau) $\sim$ 395 [31] and Taylor-Green vortex flows at Reynolds 1600, 3000 and 5000 [20]. For all these flows the modeling under uncertainty has provided better results than classical large eddies simulation models. Within the PhD of Yacine Ben Ali we will explore with the CSTB Nantes centre the application of such models for the definition of Reynolds average simulation (RANS) models for wind engineering applications.

7.3.3. Singular and regular solutions to the Navier-Stokes equations (NSE) and relative turbulent models

Participants: Roger Lewandowski, Etienne Mémin, Benoit Pinier.

The common thread of this work is the problem set by J. Leray in 1934 : does a regular solution of the Navier-Stokes equations (NSE) with a smooth initial data develop a singularity in finite time, what is the precise structure of a global weak solution to the Navier-Stokes equations, and are we able to prove any uniqueness result of such a solution. This is a very hard problem for which there is for the moment no answer. Nevertheless, this question leads us to reconsider the theory of Leray for the study of the Navier-Stokes equations in the

whole space with an additional eddy viscosity term that models the Reynolds stress in the context of large-scale flow modelling. It appears that Leray's theory cannot be generalized turnkey for this problem, so that things must be reconsidered from the beginning. This problem is approached by a regularization process using mollifiers, and particular attention must be paid to the eddy viscosity term. For this regularized problem and when the eddy viscosity has enough regularity, we have been able to prove the existence of a global unique solution that is of class C^∞ in time and space and that satisfies the energy balance. Moreover, when the eddy viscosity is of compact support in space, uniformly in time, we recently shown that this solution converges to a turbulent solution to the corresponding Navier-Stokes equations when, the regularizing parameter goes to 0. These results are described in a paper that has been submitted to the journal Archive for Rational Mechanics and Analysis (ARMA).

In the same direction, we also finalized a paper in collaboration with L. Berselli (Univ. Pisa, Italy) about the well known Bardina's turbulent model. In this problem, we consider the Helmholtz filter usually used within the framework of Large Eddy Simulation. We carry out a similar analysis, by showing in particular that no singularity occurs for Bardina's model.

Another study in collaboration with B. Pinier, P. Chandramouli and E. Memin has been undertaken. This work takes place within the context of the PhD work of B. Pinier. We considered the standard turbulent models involving the Navier-Stokes equations with an eddy viscosity that depends on the Turbulent Kinetic Energy (TKE), coupled with a supplementary equation for the TKE. The problem holds in a 3D bounded domain, with the Manning law at the boundary for the velocity. We have modeled a flux condition at the boundary for the TKE. We prove that with these boundary conditions, the resulting problem has a distributional solution. Then a serie of numerical tests has been performed in a parallelepiped with a non trivial bottom, showing the accuracy of the model in comparison with a direct numerical simulation of the Navier-Stokes equations. This work will be submitted to a journal.

7.3.4. 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 studied the impact of the introduction of a random modeling of the friction velocity on the classical wall law expression. This model derived within the modeling under location uncertainty principle enabled us to propose an improved model of the velocity profile with a clear formulation in particular in the buffer turbulent area between the viscous zone and the turbulent region. Preliminary results on chanel flows are very promising. We are actually assessing this model on turbulent boundary layer flow at high Reynold.

7.3.5. 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.

7.3.6. 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. Then this data will be used to assess the performances of the 4DVar assimilation technique developed in the context of Pranav Chandramouli's PhD to reconstruct three-dimensional turbulent flows.

7.4. Visual servoing approach for fluid flow control

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

Participants: Christophe Collewet, Johan Carlier.

This study aims at controlling 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. Our goal is to maintain the shear-layer in a desired state and thus to reject upstream perturbations. In our conference IFAC paper (https://hal.inria.fr/hal-01514361) we focused on perturbations belonging to the same space that the actuators, concretely that means that we were only able to face perturbations of the actuator itself, like failures of the actuator. This year we enlarged this result to purely exogenous perturbations. An optimal control law has been derived to minimize the influence of the perturbation on the flow. To do that, an on-line estimation of the perturbation has been used. This work will be submitted to the upcoming IEEE Conference on Decision and Control. We have also generalized the works initiated during the post-doctoral stay of Tudor-Bogdan Airimitoaie (https://hal.archives-ouvertes.fr/hal-01101089) concerning the benefits of increasing the controlled degrees of freedom in the particular case of the heat equation. This work has been validated on the shear flow.

7.5. Coupled models in hydrogeology

7.5.1. Coupling of subsurface and seepage flows

Participants: Jocelyne Erhel, Jean-Raynald de Dreuzy.

Hillslope response to precipitations is characterized by sharp transitions from purely subsurface flow dynamics to simultaneous surface and subsurface flows. Locally, the transition between these two regimes is triggered by soil saturation. Here we develop an integrative approach to simultaneously solve the sub- surface flow, locate the potential fully saturated areas and deduce the generated saturation excess over- land flow. This approach combines the different dynamics and transitions in a single partition formulation using discontinuous functions. We propose to regularize the system of partial differential equations and to use classic spatial and temporal discretization schemes. We illustrate our methodology on the 1D hillslope storage Boussinesq equations. We first validate the numerical scheme on previous numerical experiments without saturation excess overland flow. Then we apply our model to a test case with dynamic transitions from purely subsurface flow dynamics to simultaneous surface and subsurface flows. Our results show that discretization respects mass balance both locally and globally, converges when the mesh or time step are refined. Moreover the regularization parameter can be taken small enough to ensure accuracy without suffering of numerical artefacts. Applied to some hundreds of realistic hillslope cases taken from Western side of France (Brittany), the developed method appears to be robust and efficient. This study performed within the H2MNO4 ANR project has been published in the journal Advances in Water Ressources [21].

7.5.2. Characterizations of Solutions in Geochemistry

Participant: Jocelyne Erhel.

We study the properties of a geochemical model involving aqueous and precipitation-dissolution reactions at a local equilibrium. By reformulating the model as an equivalent optimization problem, we prove existence and uniqueness of a solution. It is classical in thermodynamic to compute diagrams representing the phases of the system. We introduce here the new precipitation diagram that describes the mineral speciation in function of the parameters of the system. Using the polynomial structure of the problem, we provide characterizations and an algorithm to compute the precipitation diagram. Numerical computations on some examples illustrate this approach. This work, is part of the H2MNO4 initiative. It has been recently submitted to a journal for publication [45].

7.5.3. Reactive transport in fractured-porous media

Participants: Yvan Crenner, Jean-Raynald de Dreuzy, Jocelyne Erhel.

Fractures must be carefully considered for the geological disposal of radioactive wastes. They critically enhance diffusivity, speed up solute transport, extend mixing fronts, and in turn, modify the physico-chemical conditions of reactivity in the Excavation Damaged Zone (EDZ) of the galleries. On the other hand, the pyrite oxidation could be considered like the main reaction due to the diffusion of oxygen through the gallery. Moreover, we assume that this reaction is complete in these geological conditions. First, we propose a numerical explicit reactive transport model in a fractured medium for an overall reaction. Afterwards, we present simulations outputs of the pyrite-oxygen reaction in EDZ zone. This study supported by ANDRA has been published in a conference [27].

7.5.4. Reactive transfers for multi-phasic flows

Participants: Jocelyne Erhel, Bastien Hamlat.

This study focuses on the mathematical modeling of reactive transfers for multi-phasic flows in porous medium. This study supported by IFPEN has been presented in a conference paper [37].

7.6. Linear solvers

7.6.1. Variable s-step GMRES

Participants: Jocelyne Erhel, David Imberti.

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. We introduce a new variation on s-step GMRES in order to improve its stability, reduce the number of iterations necessary to ensure convergence, and thereby improve parallel performance. In doing so, we develop a new block variant that allows us to express the stability difficulties in s-step GMRES more fully. This work supported by the EoCoE grant has been published in a conference proceeding [38] and in the journal [28].

7.6.2. Krylov method applied to reactive transport

Participant: Jocelyne Erhel.

Reactive transport models couple advection dispersion equations with chemistry equations. If the reactions are at thermodynamic equilibrium, then the system is a set of partial differential and algebraic equations. After space and implicit time discretizations, a nonlinear system of equations must be solved at each time step. The Jacobian matrix of the nonlinear system can be written with a Kronecker product coupling transport and chemistry. Krylov methods are well-suited to solve such linear systems because the matrix vector product can be done efficiently. The main challenge is to design a preconditioning matrix. We propose here to use the special structure of the matrix. Preliminary experiments show that Krylov methods are much more efficient than a direct method which does not use the coupled structure. This work supported by ANDRA has been published at the occasion of an invited conference [28].

GENSCALE Project-Team

7. New Results

7.1. Data Structure

7.1.1. Minimal perfect hash function

Participants: Antoine Limasset, Guillaume Rizk, Pierre Peterlongo.

Minimal perfect hash functions are fundamental objects used in many applications. Existing algorithms and implementations that build such functions have in practice some upper bounds on the number of input elements they can handle, due to high construction time and/or memory usage. We propose a simple algorithm having very competitive construction times, memory usage and query times compared to state of the art techniques [27]. We provide a parallel implementation called BBHash. It is capable of creating a minimal perfect hash function of 10^{10} elements in less than 1 hour and 4 GB of memory. To the best of our knowledge, this library is also the first that has been successfully tested on 10^{12} input elements. Source code: https://github.com/rizkg/BBHash

7.1.2. Quasi-dictionary

Participants: Camille Marchet, Antoine Limasset, Pierre Peterlongo.

Indexing massive data sets is extremely expensive for large scale problems. In many fields, huge amounts of data are currently generated, however extracting meaningful information from voluminous data sets, such as computing similarity between elements, is far from being trivial. It remains nonetheless a fundamental need. In this context, we proposed a probabilistic data structure based on a minimal perfect hash function for indexing large sets of keys. This structure out-competes the hash table for construction, query times and for memory usage, in the case of the indexation of a static set. To illustrate the impact of algorithms performances, we provided two applications based on similarity computation between collections of sequences, and for which this calculation is an expensive but required operation. In particular, we showed a practical case in which other bioinformatics tools failed to scale up the tested data set or provide lower recall quality results [43].

7.2. Algorithms & Methods

7.2.1. Short Read Correction

Participants: Antoine Limasset, Pierre Peterlongo.

We proposed a new method to correct short reads using de Bruijn graphs, and we implemented it as a tool called Bcool. As a first step, Bcool constructs a corrected compacted de Bruijn graph from the reads. This graph is then used as a reference and the reads are corrected according to their mapping on the graph. We showed that this approach yields a better correction than $k$mer-spectrum techniques, while being scalable, making it possible to apply it to human-size genomic datasets and beyond [41].

7.2.2. Long transcriptomic read clustering

Participants: Camille Marchet, Pierre Peterlongo.

This contribution tackles the problem of clustering RNA reads in clusters representing all variants of each gene, in a *de novo* way i.e. without any reference sequences. Such problem is not new as is, but the latest, Third Generation Sequencing (TGS) data redefine it. Reads can now span full-length transcripts but at the price of very high error rates, mostly insertions and deletions. This makes difficult or impossible to use tools designed for previous sequencing data. Still, the property to obtain whole RNA molecules through reads is very promising to better describe a transcriptome. In this work, we targeted the need to extract relevant information from a TGS transcriptome, even when no reference is available. In collaboration with Jacques Nicolas from the Inria/IRISA Dyliss team, we propose a novel algorithm in the community detection framework, based on the clustering coefficient. In addition we propose an implementation of this algorithm in the tool CARNAC-LR and a pipeline for the processing of transcriptome data. We validated our tool on real data from mouse and showed that it could be accurate and precise even for lowly expressed genes. We showed that our approach can be complementary to a mapping in the case a reference exists, and that a straightforward use of CARNAC-LR enables to quickly assess the genes'e expression levels [42].

7.2.3. Statistically Significant Discriminative Patterns Search

Participants: Hoang Son Pham, Dominique Lavenier.

Identifying multiple SNPs combinations associated with diseases such as cancers or diabetes is a central goal of human genetics. Recently, discriminative pattern mining algorithms have been investigated to tackle genome-wide association studies (GWAS). We designed an algorithm, called SSDPS, to discover groups of items which have significant difference of frequency in case-control datasets. The algorithm directly uses relative risk measures such as risk ratio, odds ratio and absolute risk reduction combined with confidence intervals as anti-monotonic properties to efficiently prune the search space. The algorithm discovers a complete set of discriminative patterns with regard to given thresholds or applies heuristic strategies to extract the largest statistically significant discriminative patterns in a given dataset. Experimental results on both synthetic datasets and three real variant datasets (Age-Related Macular Degeneration, Breast Cancer and Type 2 Diabetes) demonstrate that the SSDPS algorithm effectively detects multiple SNPs combinations in an acceptable execution time.

7.2.4. Reference free SNP detection in RAD-seq data

Participants: Jeremy Gauthier, Claire Lemaitre, Pierre Peterlongo.

We developed an original method for reference-free variant calling from Restriction site associated DNA Sequencing (RAD-Seq) data. RAD-seq is a technique widely employed in the evolutionary biology field. Based on the variant caller DiscoSnp, DiscoSnp-RAD explores the De Bruijn Graph built from all the read datasets to detect SNP and Indels. Tested on simulated and real datasets, DiscoSnp-RAD identifies thousands of variants suitable for different population genomics analyses. Furthermore, DiscoSnp-RAD stands out from other tools due to his completely different principle, making it significantly faster, in particular on large datasets [39].

7.2.5. Global Optimization for Scaffolding and Completing Genome Assemblies

Participants: Sebastien Francois, Rumen Andonov, Dominique Lavenier.

We developed a method for solving genome scaffolding as a problem of finding the longest simple path in a graph defined by the contigs that satisfies a maximal number of additional constraints encoding the insert-size information [26]. Then we solved the resulting mixed integer linear program to optimality using the Gurobi solver. We tested our algorithm on a benchmark of chloroplast genomes and showed that it outperforms other widely-used assembly solvers by the accuracy of the results.

7.2.6. Identification and characterization of long non-coding RNA

Participant: Fabrice Legeai.

We participated in the development and validation of the tool FeelNC (collaboration with IGDR group). This is a tool allowing the identification of long non coding RNA (lncRNA) from RNASeq reads with or without a reference genome. Contrary to other tools, it does not depend on the comparison with protein databanks, which usually require lots of computations, but used a machine learning approach based on a Random Forest model trained with general features such as multi k-mer frequencies and relaxed open reading frames. We delivered a module that allows to characterize the relationships of each long non coding RNA with the other genes in its genomics close environment, giving insights about the putative impact of the lncRNAs to the regulation of these genes [23], [24].

7.2.7. Characterizing repeat-associated subgraphs in de Bruijn graphs

Participant: Camille Marchet.

The main problem in genome assembly, namely repeats, is also present in transcriptomic data. They are dealt with using various heuristics in the de Bruijn Graph framework (dBG). In this work, we introduce a formal model for representing high copy-number and low-divergence repeats in RNA-seq data in dBG and infer the definition of repeat-associated subgraphs. We show that the problem of identifying such subgraphs in a dBG is NP-complete. Then we place ourselves in the case of local assembly of alternative splicing and show that such subgraphs can be avoided implicitly. Thus, more alternative splicing events can be enumerated than with previous approaches. Finally we show that this exploration of DBG explorations can improve de novo transcriptome evaluation methods [16].

7.3. Parallelism

7.3.1. Variant detection using processing-in-memory technology

Participants: Charles Deltel, Dominique Lavenier.

The concept of Processing-In-Memory aims to dispatch the computer power near the data. Together with the UPMEM company (http://www.upmem.com/), which is currently developing a DRAM memory enhanced with computing units, we investigate the parallelization of the detection of mutations on the human genome. Traditionnaly, this process is split into 2 steps: a mapping step and a variant calling step. Here, thanks to the high processing power of this new type of memory, the mapping step can nearly be done at the disk transfer rate, allowing the variant calling step to be done simultaneously on the host processor. The implementation is currently on going. First performance evaluations indicate speed-up of one or two order of magnitude compared to purely software implementation.

7.4. Bioinformatics Analysis

7.4.1. Study of marine plankton holobionts

Participants: Camille Marchet, Pierre Peterlongo.

We derived from the quasi-dictionary (described in previous section) a tool called Short Read Connector (SRC), able to find pairs of similar reads intra or inter read sets. We used SRC in meta-transcriptomics context to identify the actors of a symbiosis and help the assembly [44], [31]. The framework is the study of marina holobionts (host and its community of symbionts) for which few is known about the actors. In order to retrieve the functions that characterize such holobionts, RNA-seq reads from the sequencing of the whole holobiont are assembled *de novo*. Such assembly is prone to produce chimeras. Thus SRC is used to index sequences (reads, EST, assembled genes...) known to be close to the host and symbionts of the holobiont. Then, thanks to SRC's ability to find similarity between sequences even at a large scale, by querying reads of the holobiont we identify those similar to the host or symbionts. We report four categories: host, symbiont, shared and unassigned that can be assembled in a parallel way. As a first step we validate the SRC+assembly approach by comparing our result to literature with two known holobionts with eukaryote hosts (*Orbicella faveolata, Xestospongia muta*). We show that our approach can compare to previous results. In a second step we lean on a protist (Collodaria) holobiont for which the actors are poorly known. No assembled sequences exist in the literature so we compare the pipeline SRC+assembly to a sole assembly pipeline. Our main achievement is to highly reduce (up to $\sim$40%) the number of chimeras in the assembly compared to the sole assembly pipeline.

7.4.2. Pea aphid metagenomics

Participants: Cervin Guyomar, Fabrice Legeai, Claire Lemaitre.

We worked on a framework adapted to the study of genomic diversity and evolutionary dynamics of the pea aphid symbiotic community from an extensive set of metagenomics datasets. The framework is based on mapping to reference genomes and whole genome SNP-calling. We explored the genotypic diversity associated to the different symbionts of the pea aphid at several scales : across host biotypes, amongst individuals of the same biotype, or within individual aphids. Thorough phylogenomic analyses highlighted that the evolutionary dynamics of symbiotic associations strongly varied depending on the symbiont, reflecting different histories and possible constraints [40], [30].

7.4.3. Assembly and comparison of two genomes of highly polyphagous lepidopteran pests

Participants: Fabrice Legeai, Claire Lemaitre.

In this study, two genomes of an agronomical important lepidopteran pest, the noctuid moth *Spodoptera frugiperda*, were sequenced and compared, giving significant insights to the mechanisms involved in host-plant adaptation and speciation of this organism. In particular, we described the large expansion of gustatory receptors and detoxification genes among this polyphagous pest compared to other specialist Lepidoptera, and emphasizes the role of these 2 gene families in the evolution of one of the world's worst agricultural pests. We also provided the genome assemblies, gene annotations and whole genome alignments of both strains, and the comparison of both to a reference moth genome (*Bombyx mori*). For these purposes, several original methods were developed i) to correct genome assembly errors due to the high level of heterozygosity and ii) to extract structural variant calls from whole genome alignments [15].

7.4.4. Benchmark of de novo read dataset compression tools

Participants: Gaetan Benoit, Dominique Lavenier, Claire Lemaitre.

In this book chapter, we review the different approaches and their tools developed so far to compress sequencing data files. We detail the algorithms for each of the three main types of data contained in such files for each read : the header, the DNA and the quality sequences. We also provide a thorough benchmark of the numerous available tools on various sequencing datasets, evaluating the compression ratio as well as the running time and memory usage performances [33].

7.4.5. Genomics of the agro-ecosystems pests

Participants: Fabrice Legeai, Claire 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 or participated to the assembly and the annotation of 4 new aphids [17], [22], and 5 parasitic wasps. 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). Moreover our engagement in the agronomical pest genomics led to our contribution to other projects such as epigenetics and chromatin structure analysis [18], or the analysis of population genetics data for identifying hotspots of selection in the nematode *Globodera pallida* genome [14].

7.4.6. Comparison of approaches for finding alternative splicing events in RNA-seq

Participant: Camille Marchet.

In this work we compared an assembly-first and a mapping-first approach to analyze RNA-seq data and find alternative splicing (AS) events. Assembly-first approach enables to identify novel AS events and to detect events in paralog genes that are hard to find using mapping because of the multi-mapping results. On the other hand, the mapping-first approach is more sensitive and detects AS events in lowly expressed genes, and is also able to find AS events with exons containing transposable elements. In addition we support these results with experimental validation. We showed that in order to extensively study the alternative splicing via RNA-seq data and retrieve the most candidates, both approaches should be led. We provide a pipeline consituted of parallel local *de novo* assembly executed by KisSplice and mapping using a novel mapping workflow called FaRLine [37].

7.4.7. Microbial communities interaction between plant and their bioagressors

Participants: Susete Alves Carvalho, Fabrice Legeai, Claire Lemaitre, Pierre Peterlongo, Dominique Lavenier.

GenScale actively collaborates with the INRA group 'plant-microbial communities interactions' (IGEPP, Rennes) that analyze the interaction between plant, their associated microbial communities and different bioagressors. The ambition of the project is to understand the link between the taxonomic biodiversity of the microbiota and their functional diversity in relation with plant physiology and plant-bioagressors interactions. For this last point, an integrated metatranscritomic approach is developped. Beside wet lab and sequence productions, bioinformatics tools are needed and meta-transcriptomic pipelines analysis arecurrently developped based on the GenScale expertise.

7.5. Challenges

7.5.1. Participation to CAMI: de-novo metagenomics assembly competition

Participants: Charles Deltel, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

In metagenome analysis, computational methods for assembly, taxonomic profiling and binning are key components facilitating downstream biological data interpretation. However, a lack of consensus about benchmarking datasets and evaluation metrics complicates proper performance assessment. In this context, we participated to CAMI (Critical Assessment of Metagenome Interpretation), specifically on the assembly section with the Minia pipeline. The CAMI challenge aimed to benchmark programs on datasets of unprecedented complexity and realism. Benchmark metagenomes were generated from 700 newly sequenced microorganisms and 600 novel viruses and plasmids, including genomes with varying degrees of relatedness to each other and to publicly available ones and representing common experimental setups. Across all datasets, our assembly programs performed well for species represented by individual genomes, while performance was substantially affected by the presence of related strains [20].

SERPICO Project-Team

7. New Results

7.1. Statistical methods for image denoising and reconstruction

Participants: Emmanuel Moebel, Charles Kervrann.

In the line of the Non-Local (NL) means [39] and ND-SAFIR [11], [12], [6] denoising algorithms, we have proposed a novel adaptive estimator based on the weighted average of observations taken in a neighborhood with weights depending on image data. The idea is to compute adaptive weights that best minimize an upper bound of the pointwise L_2 risk. In the framework of adaptive estimation, we show that the "oracle" weights depend on the unknown image and are optimal if we consider triangular kernels instead of the commonly-used Gaussian kernel. Furthermore, we propose a way to automatically choose the spatially varying smoothing parameter for adaptive denoising. Under conventional minimal regularity conditions, the obtained estimator converges at the usual optimal rate. The implementation of the proposed algorithm is also straightforward. The simulations show first that our algorithm improves significantly the classical NL-means. Second, the simulations demonstrate that it is competitive when compared to state-of-the-art denoisers both in terms of PSNR values and visual quality.

Meanwhile, we investigated statistical aggregation methods which optimally combine several estimators to produce a boosted solution [13]. This approach has been especially investigated to restore spectral information in the missing wedge (MW) in cryo-electron tomography (CET). The MW is known to be responsible for several types of imaging artifacts, and arises because of limited angle tomography: it is observable in the Fourier domain and is depicted by a region where Fourier coefficient values are unknown (see Fig. 3). The proposed stochastic method tackles the restoration problem by filling up the MW by iterating following steps: adding noise into the MW (step 1) and applying a denoising algorithm (step 2). The role of the first step is to propose candidates for the missing Fourier coefficients and the second step acts as a regularizer. A constraint is added in the spectral domain by imposing the known Fourier coefficients to be unchanged through iterations. Different denoising algorithms (BM3D, NL-Bayes, NL-means...) have been compared. Furthermore, different transforms have been tested in order to apply the constraint (Fourier transform, Cosine transform, pseudo-polar Fourier transform). Finally, we showed that this strategy can be embedded into a Monte-Carlo simulation framework and amounts to computing an aggregated estimator [13]. Convincing results have been achieved (see Fig. 3) using the Fourier Shell Correlation (FSC) as an evaluation metric.

References: [18]

Collaborators: Qiyu Jin (School of Mathematical Science, Inner Mongolia University, China),
Ion Grama and Quansheng Liu (University of Bretagne-Sud, Vannes),
Damien Larivière (Fondation Fourmentin-Guilbert),
Julio Ortiz (Max-Planck Institute, Martinsried, Germany).

7.2. Algorithms for dejittering and deconvolving fluorescence Tissue MicroArray (TMA) images

Participant: Charles Kervrann.

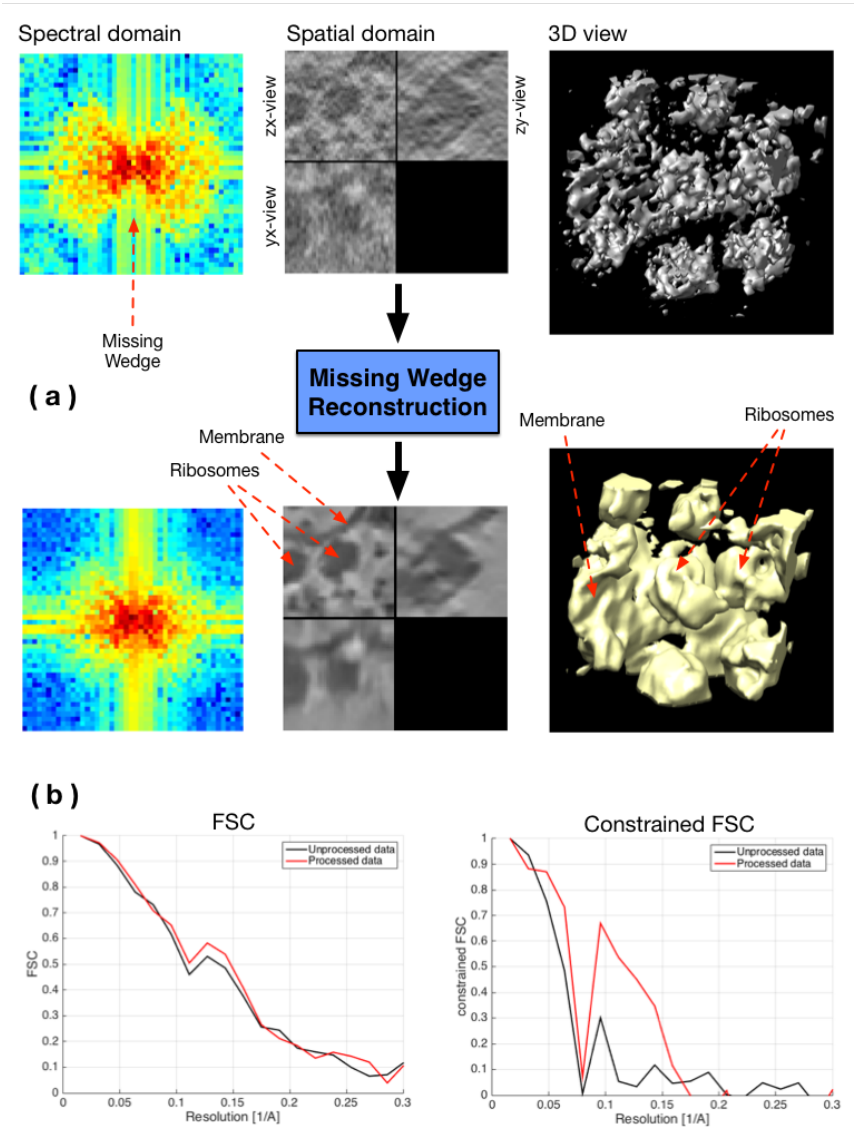

Figure 3. Experimental sub-tomogram containing ribosomes attached to a membrane. (a) Top row: original data in the spectral (left) and spatial (middle) domains and 3D view of the thresholded data (right). Bottom row: denoised data shown as above. (b) FSC and constrained FSC measures of the method input (in black) and output (in red). All measures are wrt the same reference.

In the thesis of H.-N. Nguyen, we developed dedicated image processing methods to improve quality of Tissue Microarray (TMA) images acquired by fluorescence scanners. Images are first acquired pixel by pixel along each line, with a change of scan direction between two subsequent lines. Such scanning system often suffers from pixel mis-positioning (jitter) due to imperfect synchronization of mechanical and electronic components. To correct these scanning artifacts, we proposed a variational method based on the estimation of pixel displacements on subsequent lines. This method, inspired from optical flow methods, consists in estimating a dense displacement field by minimizing an energy function composed of a non-convex data fidelity term and a convex regularization term. We used half-quadratic splitting technique to decouple the original problem into two small sub-problems: one is convex and can be solved by standard optimization algorithm, the other is non-convex but can be solved by a complete search. We showed that our method is able to remove efficiently the rolling effect due to jitter, even in the case of huge images and large non-integer displacements.

Second, to improve the resolution of acquired fluorescence images, we introduced a method of image deconvolution by considering a family of convex regularizers. The considered regularizers are generalized from the concept of Sparse Variation which combines the L1 norm and Total Variation (TV) to favors the colocalization of high-intensity pixels and high-magnitude gradient. The experiments showed that the proposed regularization approach produces competitive deconvolution results on fluorescence images, compared to those obtained with other approaches such as TV or the Schatten norm of Hessian matrix. The final deconvolution algorithm has been dedicated to large 2D 20000×60000 images acquired with ISO scan imager (see Fig.4). The method is able to process a 512×512 image in 250 ms (Matlab) with a non optimized implementation.

References: [32], [34]

Collaborators: Vincent Paveau and Cyril Cauchois (Innopys company),
Hoai-Nam Nguyen.

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

Participants: Ancageorgiana Caranfil, 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 image acquisition modalities exist, including TIRFM (Total Internal Reflection Fluorescence Microscopy), that have 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 general model of molecular interaction dynamics cannot be determined.

In the PhD thesis of A. Caranfil, we have developed a computational approach to adapt the popular temporal image correlation spectroscopy (TICS) method to the analysis of a single fusing vesicle. The biophysical diffusion model parameters (for TfR protein) are estimated by an Approximate Bayesian Computing procedure which supplies the conditional expectation and maximum a posteriori estimators from temporal correlation data. Unlike TICS, our approach is robust to noise, estimation window size, spot location and non-uniform background. It can serve in biological studies investigating diffusion processes involved in exocytosis mechanisms.

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

7.4. Classification of diffusion dynamics from particle trajectories

Participants: Vincent Briane, Charles Kervrann.

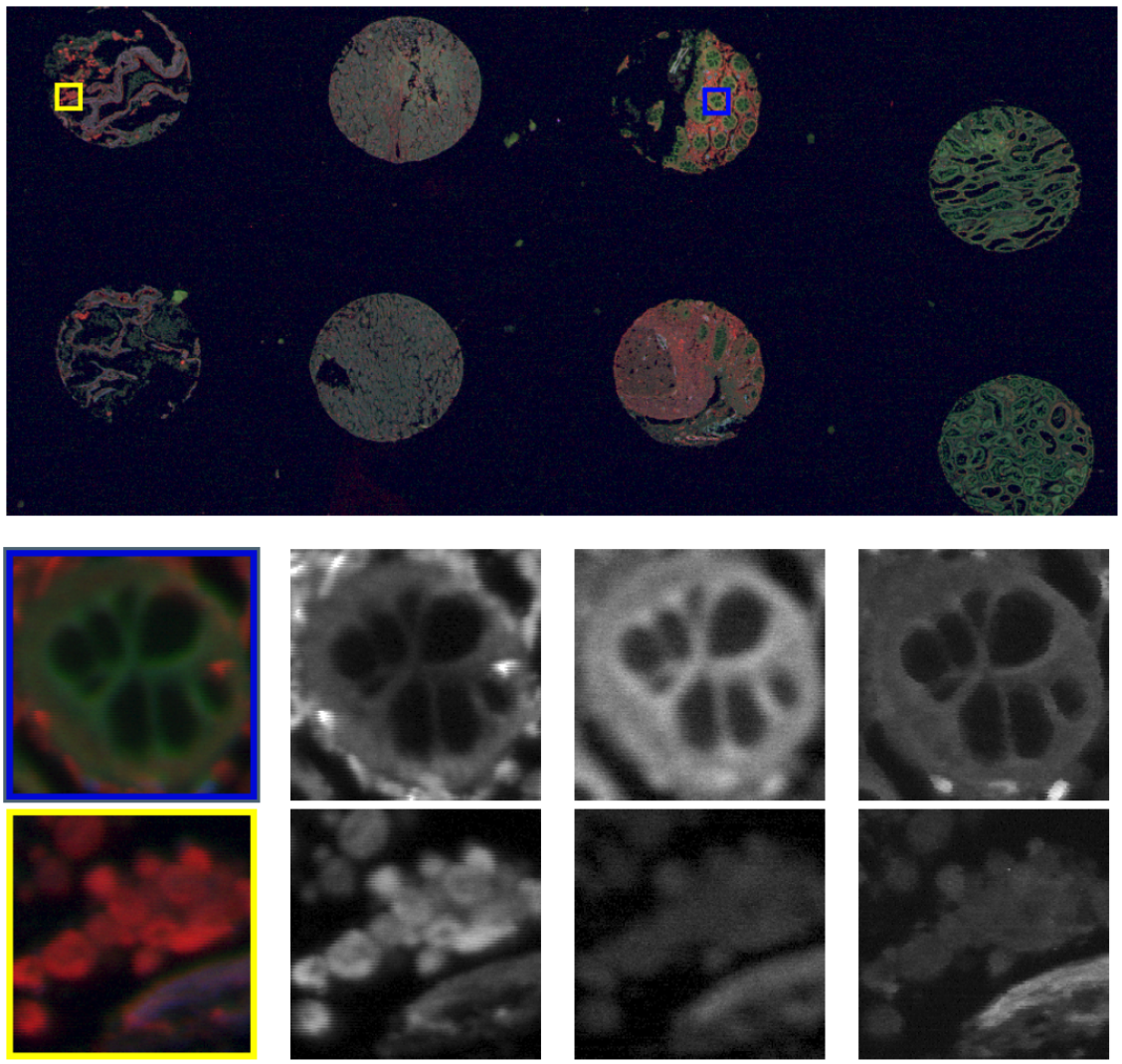

Figure 4. Three-color fluorescence image of eight tissue microarray cores. A region of interest of $4.7 \times 2.8mm^2$ *was scanned using the fluorescence scanner named InnoScan 1100AL equipped with three excitation wavelengths* ($488nm, 532nm$ *and* $635nm$) *at the spatial resolution* $0.5\mu m^2$*/ pixel, corresponding to an image of* 9544×4704 *pixels. Two areas which are bordered by two blue and yellow boxes are selected for visual comparison. First row: full size image. Second and third rows: zoom-in views of two selected areas; from left to right: 3 synchronized colors (red* ($488nm$)*, green* ($532nm$) *and blue* ($635nm$) *channels) displayed separately (courtesy of Innopsys).*

In this study, we are currently interested in describing the dynamics of particles inside live cell. Inference on the modes of mobility of molecules is central in cell biology since it reflects the interactions between the structures of the cell. In this work, we assume that the motions of particles follow a certain class of random process: the diffusion processes. Diffusions are stochastic processes with continuous paths and can model a large range of intracellular movements. Biophysicists distinguish three main types of diffusions, namely Brownian motion, superdiffusion and subdiffusion. These different diffusion processes correspond to distinct biological scenarios. A particle evolving freely inside the cytosol or along the plasma membrane is modelled by Brownian motion; the particle does not travel along any particular direction and can take a very long time to go to a precise area in the cell. Active intracellular transport can overcome this difficulty so that motion is faster in a given direction. In this case, particles are carried by molecular motors along microtubular filament networks and their motion is modelled with superdiffusion. Subdiffusion can be observed in two cases i/ when the particle is confined in a microdomain, ii/ when the particle is hindered by molecular crowding and encounters dynamic or fixed obstacles.

To address several issues in dynamics classification, we have developed a statistical test for classifying the observed trajectories into the three groups of diffusion of interest, namely Brownian motion, super-diffusion and subdiffusion. We have also designed an algorithm to detect the changes of dynamics along a single trajectory (see Fig. 5). We define the change points as the instants at which the particle switches from one diffusion type (Brownian motion, superdiffusion or subdiffusion) to another one. Finally, we have combined a clustering algorithm with our test procedure to identify micro domains, that is, zones where the particles are confined. Molecular interactions of great importance for the functioning of the cell take place in such areas.

Collaborators: Myriam Vimond (ENSAI Rennes),
Jean Salamero (UMR 144 CNRS-Institut Curie).

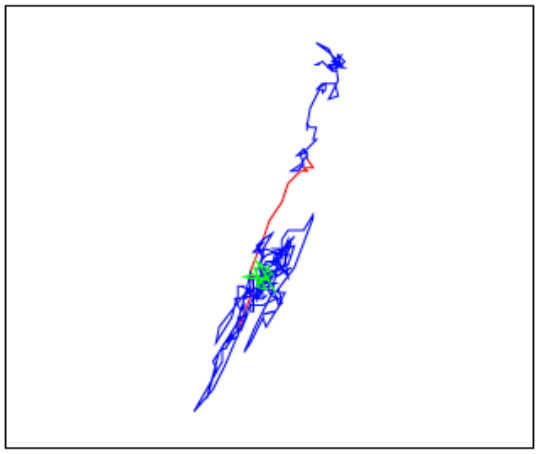

Figure 5. Change point detection on trajectories depicting neuronal mRNPs. The blue parts correspond to Brownian portions of the trajectory, red part to superdiffusive portions, green part to the subdiffusive portion.

7.5. Spatial statistics, point patterns, and colocalization in fluorescence imaging

Participants: Frédéric Lavancier, Charles Kervrann.

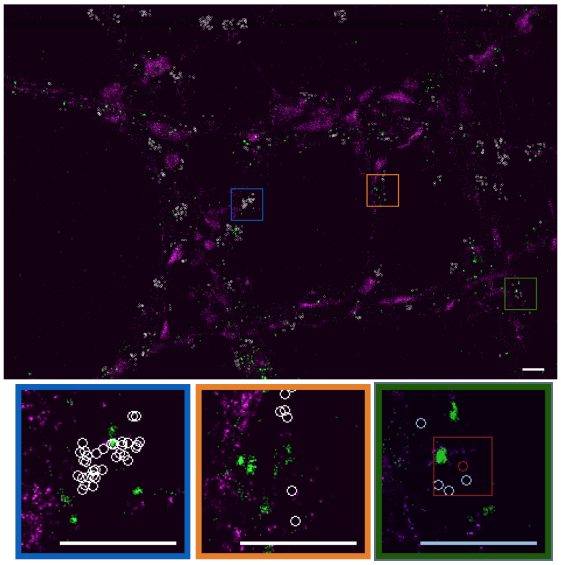

Figure 6. DSTORM acquisition of cells from hippocampi of mice expressing BDNF proteins (green channel) and vGlut (purple channel), with three zoomed-in regions (bottom). The colocalization regions identified by GcoPS are represented as white circles. The red rectangle represents the window used to find the colocalization hit shown as a red circle. The scale bars correspond to $1\mu m$

In the context of bioimaging, colocalization 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. It refers to the detection of emissions from two or more fluorescent molecules within the same pixel of the image. Colocalization is an open problem for which no satisfying solution has been found up to now. Accordingly, we proposed an objective, robust-to-noise colocalization method (GcoPS – Geo-coPositioning System)) which only requires the adjustment of a p-value that guarantees more reproducibility and more objective interpretation. It is based on the statistical analysis of the intersection (area/volume) between the two 2D or 3D binary segmented images. GcoPS handles 2D and 3D images, variable signal-to-noise ratios and any fluorescence image pair acquired with conventional or super-resolution microscopy (see Fig. 6). To our knowledge, no existing method offers the same robustness and precision level with such an easy control of the algorithm. In a recent study (internships 2017), we started to adapt this framework to analyze the spatiotemporal molecular interactions from set of 3D computed trajectories or motion vector fields (e.g., co-alignment), and then to fully quantify specific molecular machineries.

More generally, analysis of molecule and protein localization, of interactions and spatial distributions in living cells is helpful to understand functions in the cell and to compare spatialized phenotypes. This is also true with the emergence of single-molecule localization microscopy techniques (e.g., PALM), relying on the cumulative spatial localization of fluorescently tagged markers, and whose outputs are sets of spatial coordinates of single molecules. Accordingly, we were interested in the spatial distribution of single molecules that exhibit some randomness, regularity and spatial clustering (or aggregation) at large scales, while having a minimal distance between them. In that context, we theoretically studied several point processes able to represent the spatial organization of points. We focused on determinantal point processes (DDP), since they are able to describe spatial point patterns where nearby points repel or repulse each other. We also partly solved a 30 years old conjecture by proving the consistency of the likelihood procedure for a large class of Gibbs models (e.g., Strauss model, area-interaction model) which are commonly used models in practice. We extended the pseudo-

likelihood procedure to infinite range Gibbs interactions, and we proved its consistency and its asymptotic normality. All these models are now well understood and will be used in future works to analyse point patterns in cell imaging, generally described by Poisson point processes.

References: [30], [31], [35]

Collaborators: Jean Salamero and Liu Zengzhen (UMR 144 CNRS-Institut Curie),
David Dereudre (Laboratoire Paul Painlevé (UMR 8524), University of Lille 1),
Jean-François Coeurjolly (Laboratoire Jean Kutzmann, University of Grenoble).

7.6. Data assimilation and modeling of cell division mechanism

Participants: Ancageorgiana Caranfil, Charles Kervrann.

Nowadays, medical challenges demand a profound understanding of cellular mechanisms. Research in biology, biophysics and medical domain unravelled a significant part of the general processes occurring at the cellular level. It has enabled the understanding of much smaller scale processes, but our knowledge on these mechanisms is still limited as new, more complex issues need to be solved. In this context, we aim at understanding the role and interaction of the molecular key players at different scales, and their individual and collective impact on the global mechanism at the cell level. To this purpose, we have focused on the dynamics of the spindle during cell division mechanism. Our approach consists in creating a biophysical model for this mechanism, and uses data assimilation to adjust the model and optimally integrate the information from the observations. The overall spindle behaviour is led by the spindle poles behaviour. This year, we have proposed a new biophysical model for the posterior spindle pole functioning during metaphase and anaphase, that explains the oscillatory behaviour with a minimum number of parameters. Estimating the model parameters is ongoing, and will provide insights on molecular players role as well as guidance for future experiments to further investigate the dynamics of the spindle during cell division. First, we have focused on the temporal aspect. Spatial information on microtubules and molecular motors will be included in the model in the next part of this work.

Collaborators: Yann Le Cunff and Jacques Pécréaux (IGDR Institute of Genetics & Development of Rennes).

7.7. Quantifying the spatial distribution of intracellular events

Participant: Charles Kervrann.

Automated processing of fluorescence microscopy data allows to quantify cell phenotypes in an objective and reproducible way. However, most computational methods are based on the complex combination of heterogeneous features expressing geometrical, morphological and frequency properties, which makes difficult to draw definitive biological conclusions. Additionally, most experimental designs pool together data coming from replicated experiments of a given condition, neglecting the biological variability between individual cells. Hence, we developed a generic and nonparametric density framework (QuantEv) to discriminate spatiotemporal distributions (using circular Earth mover's distance) of moving proteins detected by any appropriate algorithm. The main advantage of QuantEv is to robustly process 2D and 3D data, and accurately analyse homogeneous and heterogeneous populations. As proof-of-principle, we first quantitatively characterized protein trafficking of Rab6 positive membranes between the Golgi apparatus and the plasma membrane. Next, we demonstrated that Rab11 positive membranes uniformly distribute around the Endosomal Recycling Compartment (ERC), regardless of the cell shape. Finally, we showed that actin organization is cell shape dependent, and evaluated its influence on the distribution of exocytosis/recycling vesicles. QuantEv is a flexible method which enables to quantify any intracellular trafficking in 3D flat or rounded, constrained or non-constrained, adherent or non-adherent cells.

References: [36]

Collaborators: Thierry Pécot (Hollings Cancer Center, Medical Univ. South Carolina, Charleston, USA),
Jean Salamero, Jérôme Boulanger and Liu Zengzhen (UMR 144 CNRS-Institut Curie).

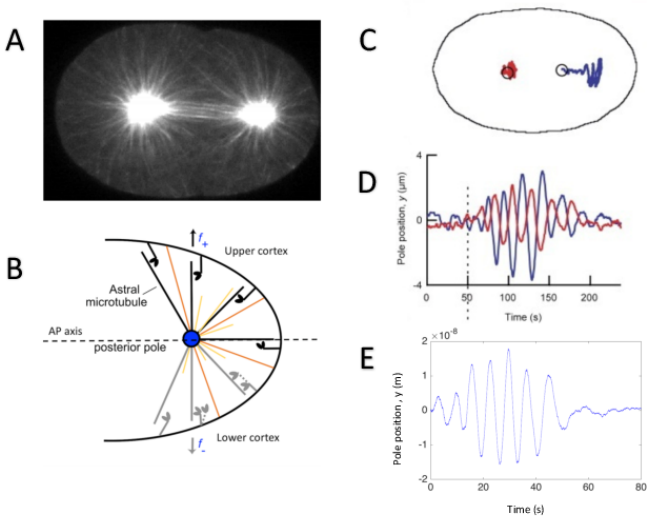

Figure 7. Illustration of the cell division mechanism observed in fluorescence microscopy (A). Sketch of one centrosome and connected to microtubules in the cell (B), experiments and tracking of the two centrosomes (C and D), and simulation of centrosome oscillations (E).

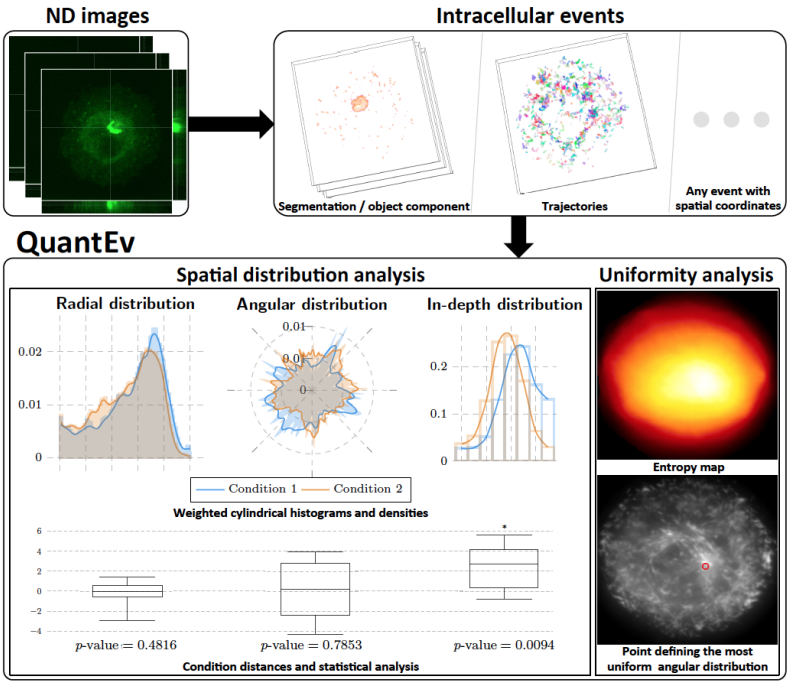

Figure 8. Overview of QuantEv approach.

7.8. 3D registration for correlative light-electron microscopy

Participants: Bertha Mayela Toledo Acosta, Patrick Bouthemy.

In recent years, correlative light and electron microscopy (CLEM) has become an attractive tool in the bio-imaging field. Biologists can collect complementary information from light microscopy (LM) and electron microscopy (EM), respectively on the dynamics and on the structure of the cell. An overlay of the LM and EM images is needed to combine information from the LM and EM sources. We are developing a 3D automated CLEM method to register EM and LM image stacks. Given the significant gap between the field of view, position and orientation of the EM and LM stacks, it is not possible to estimate directly the 3D registration. We first compute a 2D maximum intensity projection (MPI) of the LM stack along the Z-axis, and we match 2D EM regions of interest (ROI), extracted from different EM slices, into the 2D LM-MPI image. From the resulting location candidates, we estimate with a robust criterion the 2D XY-shift to pre-align the LM and EM stacks. Afterwards, a 3D affine transformation between 3D-LM-ROI and 3D-EM–ROI can be estimated using mutual information. We successfully tested this framework on two first 3D correlative microscopy datasets.

Collaborators: Xavier Heiligenstein (UMR 144 CNRS-Institut Curie),
Grégoire Malandain (Inria, Morpheme EPC, Sophia-Antipolis).

7.9. Fast optical flow methods for 3D fluorescence microscopy

Participants: Sandeep Manandhar, Patrick Bouthemy, Charles Kervrann.

Estimating motion of cells and of subcellular particles is crucial for deciphering cell mechanisms and understanding cell behaviors. Modern 3D light microscopy (LM) for cell biology enables to observe cell dynamics at a good resolution, in both space and time, motivating the development of 3D optical flow methods. However, the acquired 3D LM image sequences exhibit several specificities making 3D motion computation a difficult problem. We have defined an original and efficient two-stage estimation method for light microscopy image volumes. The method, developed in the frame of S. Manandhar PhD thesis, takes a pair of LM image volumes as input, segments the 2D slices of the source volume in super-pixels, and first estimates the 3D displacement vectors of the super-pixel centers. To this end, we have extended the well-known PatchMatch method to 3D volumes, where correspondences act between voxels. Both the propagation and the random search steps were adapted to 3D volumes. Then, a weighted interpolation has been designed to recover the dense 3D flow field for all the voxels, from the sparse 3D displacement field. The super-pixel segmentation is exploited to define the neighborhood for interpolation, and the interpolation weights take into account intensity edges and local motion differences to preserve flow discontinuities. The experimental results show good gain in execution speed, and accuracy evaluated in computer-generated 3D data with ground-truth. The results were promising on two real 3D LM image sequences supplied by USTW. The sequences depict blebbing in a MV3 cell (see Fig. 9). The cell membrane protrudes increasing the surface area of the cell. These protrusions, referred to as blebs, appear and disappear in interval of minutes, the bleb appearance corresponding to the stretching of a local region of the cell membrane. The total computation time was for the first sequence 163 seconds (resp. 101s for the second sequence), with 19 (resp. 49), 120 (resp. 44) and 24 (resp. 8) seconds for super-pixel generation, 3D patch matching, and interpolation respectively, on a computer with 2.8 GHz Intel i7 processor and 16 GB of RAM.

Collaborators: Philippe Roudot and Gaudenz Danuser (UTSW, Dallas, USA).

7.10. 3D Convolutional Neural Networks for macromolecule localization in cryo-electron tomograms of intact cells

Participants: Emmanuel Moebel, Charles Kervrann.

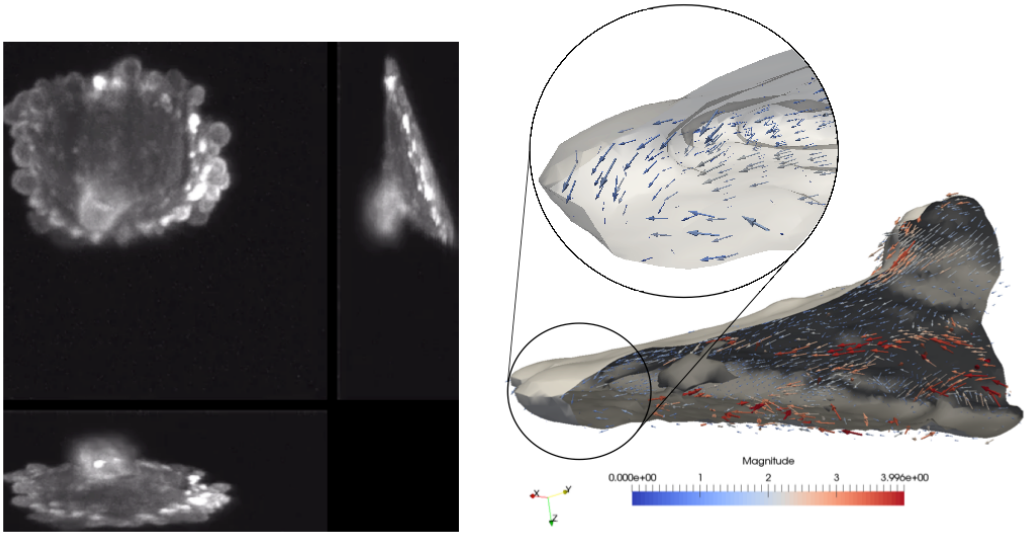

Figure 9. Illustration of 3D optical flow computation to analyze bleb deformation during cell migration in Bessel beam light sheet microscopy (input images by courtesy of Danuser lab, UTSW Dallas, USA).

In this study, we focus on macromolecule localization and classification in cryo-electron tomography (CET) images. Biologists are in need for efficient methods to localize macro-molecules (e.g. ribosomes) in frozen cell samples. The high amount of noise and imaging artifacts are the reasons why very few computational methods exist for this task. In fact, the most used method today is template matching (TM) whose resulting score map comprises a high amount of false positives. Therefore, it is necessary to apply post-processing techniques (ROI selection, classification) in order to refine the localization results. We propose an alternative localization method to TM, based on a convolutional neural network (CNN). The idea is to propose a robust and more straight-forward approach, allowing to bypass the conventional processing chain. By using python toolboxes optimized for GPU computing (elektronn, keras), we are able to reach computation time much lower than the current approach. Results on synthetic data demonstrate the superiority of our approach compared to TM. In addition, we applied our method on experimental data in order to localize sub-classes of ribosomes (membrane-bound and cytoplasmic ribosomes), a task difficult to achieve with TM alone. We are currently in the process of publishing these results. Future perspectives include localizing smaller macro-molecules, like proteasomes.

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

7.11. Estimation of parametric motion models with deep neural networks

Participants: Juan Manuel Perez Rua, Patrick Bouthemy.

We have proposed an end-to-end learning architecture for estimating a parametric motion model for a moving scene. We handle motion outliers by using the supervised training trick that is used by stacked denoising auto-encoders. Here, we define motion outliers as regions of the image whose motion does not correspond with the estimated parametric motion model. In other words, we seek to find a parametrized dominant motion of the dynamic scene. We leverage stacked hourglass networks with a final hard-coded block corresponding to

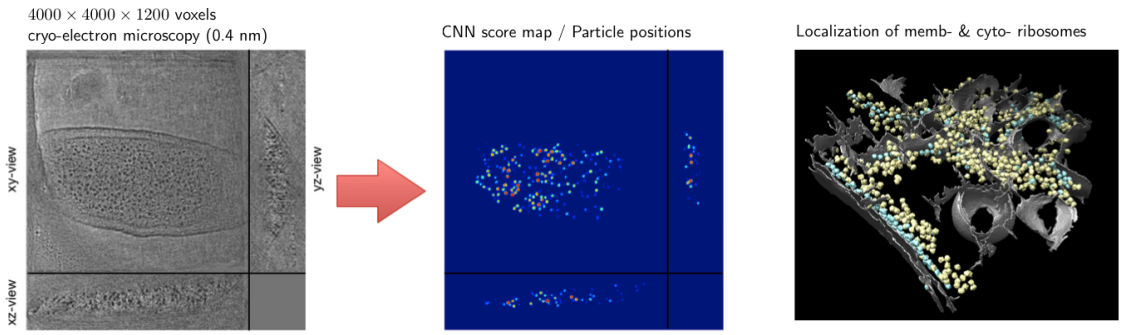

Figure 10. Illustration of 3D CNN to localize ribosomes isolated in the cytoplasm and close to the cell membrane in cryo-electron tomography (courtesy of Max-Planck Institute, Martinsried, Germany).

the global parametric motion model estimator. This block replaces the decoder part of a convolutional auto-encoder network, and it is end-to-end trainable since it involves linear operations only. Moreover, the hard-wired decoder allows the network to output the values of the parametric motion model given an input moving scene, even when the supervision acts on optical flow maps and not the motion model values. This means that our network is able to provide, as a by-product, a concise code that can be used as motion descriptor.

Collaborators: Tomas Crivelli and Patrick Pérez (Technicolor).

7.12. Motion saliency in video sequences

Participants: Léo Maczyta, Patrick Bouthemy.

Dynamic (or motion) saliency is a means to detect unexpected or rare dynamic behaviors in video sequences acquired by a stationary or a mobile imaging device. Finding salient dynamic information in each image of a sequence is indeed crucial in many situations. We aim to extract saliency only from motion information, and to exhibit salient motion in contrast to its space-time context with no prior on the nature of both. So far, we have investigated a simpler problem than saliency map estimation. We deal with the classification of each image of a sequence as dynamically salient or not, that is, containing salient motion or not. We have explored convolutional neural network (CNN). We have designed two different networks. The first one relies on two intensity images, the first input image and the second image warped with the parametric dominant motion estimated between the two input images. The second one takes as input the difference between the computed optical flow and parametric dominant flow.

Collaborators: Olivier Lemeur (EPC Sirocco, Inria Rennes - Bretagne Atlantique).

VISAGES Project-Team

7. New Results

7.1. Research axis 1: Medical Image Computing in Neuroimaging

Extraction and exploitation of complex imaging biomarkers involve an imaging processing workflow that can be quite complex. This goes from image physics and image acquisition, image processing for quality control and enhancement, image analysis for features extraction and image fusion up to the final application which intends to demonstrate the capability of the image processing workflow to issue sensitive and specific markers of a given pathology. In this context, our objectives in the recent period were directed toward 4 major methodological topics:

7.1.1. Diffusion imaging

7.1.1.1. L2 Similarity Metrics for Diffusion Multi-Compartment Model Images Registration

Participants: Renaud Hédouin, Olivier Commowick, Emmanuel Caruyer, Christian Barillot.

Diffusion multi-compartment models (MCM) allow for a fine and comprehensive study of the white matter microstructure. Non linear registration of MCM images may provide valuable information on the brain for example through population comparison. State-of-the-art MCM registration however relies on pairing-based similarity measures where the one-to-one mapping of MCM compartments is required. This approach leads to non differentiabilties or discontinuities, which may turn into poorer registration. Moreover, these measures are often specific to one MCM compartment model. We proposed [34] two new MCM similarity measures based on the space of square integrable functions, applied to MCM characteristic functions. These measures are pairing-free and agnostic to compartment types. We derived their analytic expressions for multi-tensor models and proposed a spherical approximation for more complex models. Evaluation was performed on synthetic deformations and inter-subject registration, demonstrating the robustness of the proposed measures.

7.1.1.2. Block-Matching Distortion Correction of Echo-Planar Images with Opposite Phase Encoding Directions

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

By shortening the acquisition time of MRI, Echo Planar Imaging (EPI) enables the acquisition of a large number of images in a short time, compatible with clinical constraints as required for diffusion or functional MRI. However such images are subject to large, local distortions disrupting their correspondence with the underlying anatomy. The correction of those distortions is an open problem, especially in regions where large deformations occur. We have proposed a new block-matching registration method to perform EPI distortion correction based on the acquisition of two EPI with opposite phase encoding directions (PED). It relies on new transformations between blocks adapted to the EPI distortion model, and on an adapted optimization scheme to ensure an opposite symmetric transformation. We have produced qualitative and quantitative results of the block-matching correction using different metrics on a phantom dataset and on in-vivo data. We have shown the ability of the block-matching to robustly correct EPI distortion even in strongly affected areas. This work has been published in IEEE Transactions on Medical Imaging [21].

7.1.1.3. Diffusion MRI processing for multi-compartment characterization of brain pathology

Participants: Renaud Hédouin, Olivier Commowick, Christian Barillot.

Diffusion weighted imaging (DWI) is a specific type of MRI acquisition based on the direction of diffusion of the brain water molecules. It allows, through several acquisitions, to model the brain microstructure, as white matter, which is significantly smaller than the voxel-resolution. To acquire a large number of images in a clinical setting, very-fast acquisition techniques are required as single-shot imaging. However these acquisitions suffer locally large distortions. We have proposed a block-matching registration method based on the acquisition of images with opposite phase-encoding directions (PED). This technique specially designed for Echo-Planar Images (EPI) robustly correct images and provides a deformation field. This field is applicable to an entire DWI series from only one reversed EPI allowing distortion correction with a minimal acquisition time cost. This registration algorithm has been validated both on phantom and on *in vivo* data and is available in our source medical image processing toolbox Anima. From these diffusion images, we are able to construct multi-compartments models (MCM) which can represent complex brain microstructure. Doing registration, averaging and atlas creation on these MCM images is required to perform studies and statistic analyses. We propose a general method to interpolate MCM as a simplification problem based on spectral clustering. This technique, which is adaptable for any MCM, has been validated on both synthetic and real data. Then, from a registered dataset, we performed a patient to population analysis at a voxel-level computing statistics on MCM parameters. Specifically designed tractography can also be used to make analysis, following tracks, based on individual anisotropic compartments. All these tools are designed and used on real data and contribute to the search of biomakers for brain diseases such as multiple sclerosis.

7.1.1.4. The challenge of mapping the human connectome based on diffusion tractography

Participant: Emmanuel Caruyer.

Tractography based on non-invasive diffusion imaging is central to the study of human brain connectivity. To date, the approach has not been systematically validated in ground truth studies. Based on a simulated human brain data set with ground truth tracts, we organized an open international tractography challenge, which resulted in 96 distinct submissions from 20 research groups. Here, we report the encouraging finding that most state-of-the-art algorithms produce tractograms containing 90 percent of the ground truth bundles (to at least some extent). However, the same tractograms contain many more invalid than valid bundles, and half of these invalid bundles occur systematically across research groups. Taken together, our results demonstrate and confirm fundamental ambiguities inherent in tract reconstruction based on orientation information alone, which need to be considered when interpreting tractography and connectivity results. Our approach provides a novel framework for estimating reliability of tractography and encourages innovation to address its current limitations [26].

7.1.1.5. Comparison of inhomogeneity distortion correction methods in diffusion MRI of the spinal cord

Participants: Haykel Snoussi, Emmanuel Caruyer, Christian Barillot.

Diffusion MRI (dMRI) is a modality that describes the geometry of neural architecture. Diffusion images suffer from various artifacts originating from subject and physiological motion, eddy currents and B0-field inhomogeneity. These can severely affect image quality particularly in the spine region. However, strategies exist to correct these distortions, including co-registration, point spread function, phase field map and reversed gradient polarity method (RGPM). We evalute various correction methods using RGPM which provides best results. More precisely, we compared Voss plus two other recent methods: Topup (FSL) and HySCO (ACID/SPM). This work was presented at the ESMRMB conference [38].

7.1.2. Arterial Spin Labeling:

Our contributions on this topic are illustrated in Fig. 2 . Arterial Spin Labeling (ASL) enables measuring cerebral blood flow in MRI without injection of a contrast agent. Perfusion measured by ASL carries relevant information for patients suffering from pathologies associated with singular perfusion patterns.

However this technique suffers from drawbacks such as low signal to noise ratio and poor resolution.

7.1.2.1. Patch-based super-resolution for arterial spin labeling MRI

Participants: Cédric Meurée, Pierre Maurel, Christian Barillot.

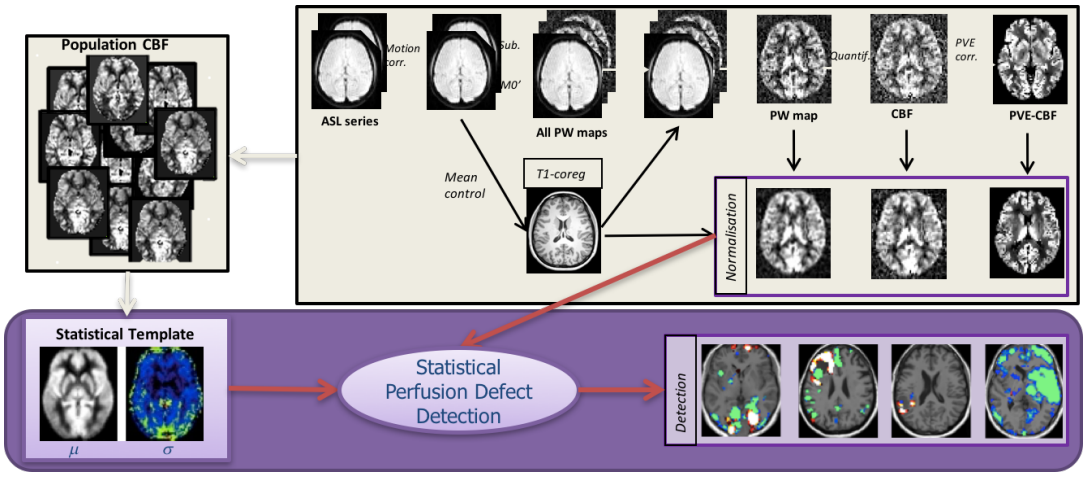

Figure 2. Summary of the image processing workflow that allows the quantification of brain perfusion and detection of potential perfusion defect on patients or populations

In this context, our contributions focused on a super resolution approach to reduce the influence of Partial Volume Effects (PVE) and obtain images close to the ones that would be acquired at a high resolution, but in a shorter scan duration. PVE are an important limitation of arterial spin labeling (ASL) acquisitions, impacting the validity of quantitative cerebral blood flow (CBF) estimations. This work consists of a super-resolution algorithm, which includes information of high resolution (HR) structural images to reconstruct HR CBF maps from low resolution ASL series, without increasing the acquisition time. Compared with nearest neighbor, trilinear and 3rd order spline interpolations, the proposed algorithm is found to generate a CBF image closer to the one obtained with a reference HR ASL acquisition. CBF calculations can therefore be improved by using this algorithm, which reduces the PVE [36].

7.1.2.2. Resting-state functional ASL

Participants: Corentin Vallée, Isabelle Corouge, Pierre Maurel, Christian Barillot.

We have started to work on resting-state functional ASL (rs-fASL). Rs-fASL in clinical daily practice and academic research stay discreet compared to resting-state BOLD. However, by giving direct access to cerebral blood flow maps, rs-fASL could lead to significant clinical subject scaled application as CBF can be considered as a biomarker in common neuropathology. As a new topic, we started by building a viable long sequence for rs-fASL. We take advantage of the long duration of the sequence to assess the link between overall quality of rs-fASL and duration of acquisition. To this end, we consider typical functional areas of the brain, and assess their quality compared to gold standards depending on the duration of acquisition. While some more work remain to be done, we tend to show there is an optimal duration of acquistion for rs-fASL. This work was submitted for the next ISMRM Conference.

7.1.2.3. Longitudinal atlas creation and brain development analysis

Participants: Antoine Legouhy, Olivier Commowick, Christian Barillot.

The study of brain development provides insights in the normal trend of brain evolution and enables early detection of abnormalities. We propose a method to quantify growth in three arbitrary orthogonal directions of the brain through linear registration. We introduce a 9 degrees of freedom transformation that gives the opportunity to extract scaling factors describing brain growth along those directions by registering a data base of subjects in a common basis. We apply this framework to create a longitudinal curve of scaling ratios along fixed orthogonal directions from 0 to 16 years highlighting anisotropic brain development. In pediatric

image analysis, the study of brain development provides insights in the normal trend of brain evolution and enables early detection of abnormalities. Tools like longitudinal atlases allow to compute statistics on populations, understand brain variability at different ages to highlight changes in growth, shape, structure etc. We experimented different methods to perform longitudinal atlases. This work was submitted for the next ISMRM Conference.

7.1.3. Quantitative relaxation times estimation and processing:

The VisAGeS team has proposed new methodologies to exploit new relaxometry sequences, able to provide direct information on tissue properties (T1, T2, T2* relaxation times) and their alteration in diseases. Such sequences have a great potential in diagnostic and evolution study of patients suffering from various neurological diseases.

7.1.3.1. Gaining Insights Into Multiple Sclerosis Lesion Characteristics from Brain Tissue Microstructure Information: A Multi-Compartment T2 Relaxometry Approach:
Participants: Sudhanya Chatterjee, Olivier Commowick, Christian Barillot.

In addition to raw relaxation times, we have also studied other estimation methods able, from T2 relaxometry sequences, to estimate the fraction of myelin (myelin water fraction) inside each voxel, a quantity that may be largely impacted in neurological diseases. To this end, we have proposed new multi-compartment T2 estimation methods [42] with a new water three-compartment T2 model of tissue bounded water (free water, axons and cells, and myelin), using variable projection to make the estimation faster and more robust. Clinical trends and pathogenetic ways of onset and progression of Multiple Sclerosis (MS) in patients suggest that MS is a highly heterogeneous disease. MS is predominantly a White Matter (WM) disease, which is mainly composed of myelinated axons and neuroglia type cells. Demyelination and axonal loss characterize the condition of MS in a patient. However, they follow varying trends in patients. In this work, we propose a method in which T2 relaxometry data is used to obtain a quantitative brain tissue microstructure information. This information is then studied to check its corroborations with pathogenetic understanding of MS in literature [41].

7.1.3.2. Multi-Compartment T2 Relaxometry Model Using Gamma Distribution Representations: A Framework for Quantitative Estimation of Brain Tissue Microstructures:
Participants: Sudhanya Chatterjee, Olivier Commowick, Christian Barillot.

Advanced MRI techniques (e.g., d-MRI, MT, relaxometry etc.) can provide quantitative information of brain tissues. Image voxels are often heterogeneous in terms of microstructure information due to physical limitations and imaging resolution. Quantitative assessment of the brain tissue microstructure can provide valuable insights into neurodegenerative diseases (e.g., Multiple Sclerosis). In this work, we propose a multicompartment model for T2-Relaxometry to obtain brain microstructure information in a quantitative framework. The proposed method allows simultaneous estimation of the model parameters [42].

7.1.4. Multi-modal EEG and fMRI Source Estimation Using Sparse Constraints:
Participants: Saman Noorzadeh, Pierre Maurel, Christian Barillot.

In this work, a multi-modal approach is presented and validated on real data to estimate the brain neuronal sources based on EEG and fMRI. Combining these two modalities can lead to source estimations with high spatio-temporal resolution. The joint method is based on the idea of linear model already presented in the literature where each of the data modalities are first modeled linearly based on the sources. Afterwards, they are integrated in a joint framework which also considers the sparsity of sources. The sources are then estimated with the proximal algorithm. The results are validated on real data and show the efficiency of the joint model compared to the uni-modal ones. We also provide a calibration solution for the system and demonstrate the effect of the parameter values for uni- and multi-modal estimations on 8 subjects [37].

7.2. Research axis 2: Applications in Neuroradiology and Neurological Disorders

7.2.1. Arterial Spin Labeling:

Participants: Jean-Christophe Ferré, Maia Proisy, Isabelle Corouge, Élise Bannier, Christian Barillot.

Arterial Spin Labeling is an attractive perfusion MRI technique due to its complete non-invasiveness. However it still remains confidential in clinical practice. Over the years, we have developed several applications to evaluate its potential in different contexts. In 2017, in the context of the MALTA project, we focused on the application of ASL to activation-fMRI. 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 was 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, we explored the ability of fASL to map the motor areas at different post-labeling delays (PLD) in healthy subjects and patient with brain tumor. As part of the PhD of Maia Proisy, we have also been working on processing and analyse MR perfusion images using arterial spin labeling in neonates and children for several purposes:

- ASL and TOF-MRA are two totally non-invasive, easy-to-use MRI sequences for children in emergency settings. Hypoperfusion associated with homolateral vasospasm may suggest a diagnosis of migraine with aura (published in Cephalagia and presented in 3 congresses including RSNA)
- Investigation of brain perfusion evolution between 6 month and 15 years using ASL sequence in order to provide reference values in this age range (Measurement of pediatric regional cerebral blood flow from 6 months to 15 years of age article under revision, presented in one national congress)
- Work in Progress: ASL perfusion images in 20 neonates with hypoxic-ischemic encephalopathy that underwent MRI on day-of-life 3 and day-of-life 10.

7.2.2. Hybrid EEG-fMRI Neurofeedback:

Participants: Lorraine Perronnet, Marsel Mano, Élise Bannier, Mathis Fleury, Giulia Lioi, Christian Barillot.

Over the last 4 years, we developed a whole new range of activities around hybrid EEG-MR imaging and neurofeedback for brain rehabilitation. We propose to combine advanced instrumental devices (Hybrid EEG and MRI platforms), with new man-machine interface paradigms (Brain computer interface and serious gaming) and new computational models (source separation, sparse representations and machine learning) to provide novel therapeutic and neuro-rehabilitation paradigms in some of the major neurological and psychiatric disorders of the developmental and the aging brain. We first performed a thorough state-of-the-art of Neurofeedback (NF) and restorative Brain Computer Interfaces (BCI) under EEG and fMRI modality as well as of EEG-fMRI integration, with a particular focus on applications in depression and motor rehabilitation. This enabled us to design a NF protocol based on motor imagery and compatible with EEG and fMRI. We implemented different types of feedback and compared for the first time the effects of unimodal EEG-NF and fMRI-NF versus bimodal EEG-fMRI-NF by looking both at EEG and fMRI activations. We also introduced a new feedback metaphor for bimodal EEG-fMRI-neurofeedback that integrates both EEG and fMRI signal in a single bi-dimensional feedback (a ball moving in 2D). The participants to this study were able to regulate activity in their motor regions in all NF conditions. Our results also suggest that that EEG-fMRI-neurofeedback could be more specific or more engaging than EEG-NF alone [31].

All the experiments were performed on the Neurinfo platform which is equipped with an EEG MR compatible 64-channel device in 2014 to perform joint EEG and BOLD or ASL fMRI. We developed, installed and successfully tested a hybrid EEG-fMRI platform for bimodal NF experiments. Our system is based on the integration and the synchronization of an MR-compatible EEG and fMRI acquisition subsystems. We developed two real-time pipelines for EEG and fMRI that handle all the necessary signal processing, the joint NF block that calculates and fuses the NF and a visualization block that displays the NF to the subject. The control and the synchronization of both subsystems with each other and with the experimental protocol

is handled by the NF Control. Our platform showed very good real-time performance with various pre-processing, filtering, and NF estimation and visualization methods. Its modular architecture is easily adaptable to different experimental environments, and offers high efficiency for optimal real-time NF applications [27].

These developments came as part of the HEMISFER project which is conducted through a very complementary set of competences over the different teams involved (Visages Inserm U1228, HYBRID and PANAMA Teams from Inria/IRISA, EA 4712 team from University of Rennes I and ATHENA team from Inria Sophia-Antipolis). The overall principle of this project is illustrated in Fig. 3 .

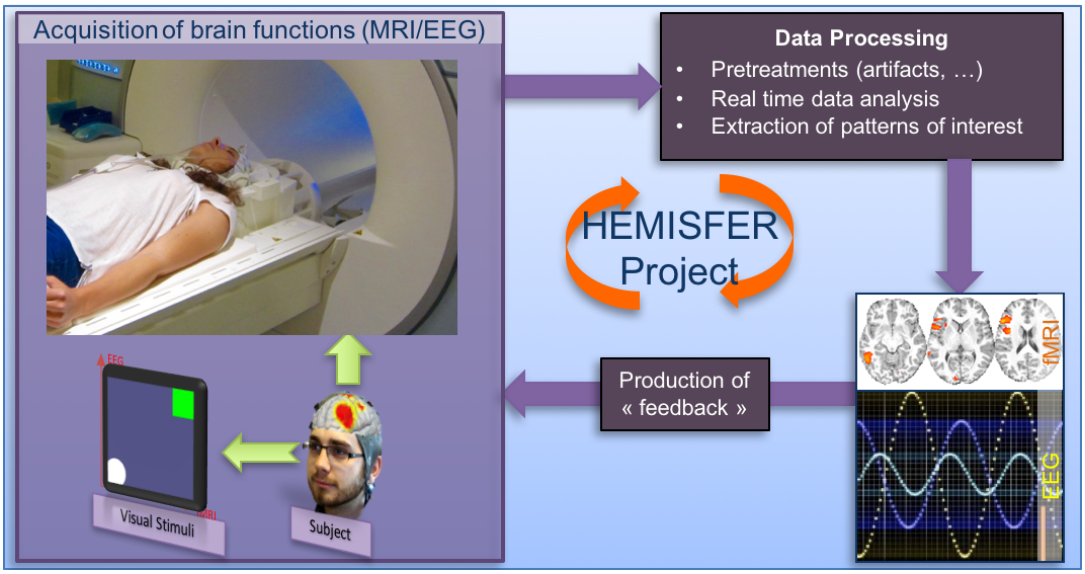

Figure 3. Principle of the Hybrid EEG:fMRI environement set up and used by the HEMISFER project

7.2.3. Multiple sclerosis:

Participants: Anne Kerbrat, Gilles Edan, Jean-Christophe Ferré, Benoit Combès, Olivier Commowick, Élise Bannier, Sudhanya Chatterjee, Haykel Snoussi, Emmanuel Caruyer, Christian Barillot.

The VisAGeS research team has a strong focus on applying the developed methodologies (illustrated in research axis 1) to multiple sclerosis (MS) understanding and the prediction of its evolution. Related to the EMISEP project on spinal cord injury evolution in MS, a first work investigated the magnetization transfer reproducibility across centers in the spinal cord and was accepted for presentation at ESMRMB [33]. Based on this work, a second work investigated the sensitivity of magnetization transfer to assess diffuse and focal burden in MS patients [43]. In parallel, methodological developments have addressed spinal cord diffusion data analysis, starting with a comparaison of several distortion correction methods [38].

Finally, we investigated myelin water fraction (MWF) estimation on multiple sclerosis and demonstrated in longitudinal studies [41] how these figures can be related with lesion evolution, paving the way towards myelin oriented MS evaluation of patient future evolution prediction (and thus treatment adaptation) and joint studies between different quantitative imaging modalities (e.g., diffusion).

7.2.4. Recovery imaging:

Participants: Isabelle Bonan, Stephanie Leplaideur, Élise Bannier, Jean-Christophe Ferré, Christian Barillot.

More common after a right hemispheric brain injury, misperception of body in space, impacting moves and posture is often associated with disturbance of spatial attention (behavioural symptoms of a failure in spontaneously reorienting attention to stimulus information in the left field). While different subjects use different references in their elaboration of spatial representation, body-centered coordinate systems are the most prevalent. As part of an fMRI substudy of a national research study on balance disorder rehabilitation, we investigated differences in activations during body-centered spatial tasks in corporeal and in extracorporeal space. Healthy controls and stroke patients were included in this fMRI sub study comprising 2 egocentric spatial tasks: perception of the midsagittal plane in extracorporeal space (straight-ahead task) and in corporeal space (longitudinal axis task). Results obtained on healthy control data were presented at the SOFMER conference and the journal paper is under review. For both tasks, cerebral activations largely dominated in the right hemisphere and essentially involved the right frontoparietal network. In addition, the straight-ahead task presented specific activations in the temporoparieto-insular cortex and thalamic areas. Patient data processing is ongoing in the context of an MD-PhD. In parallel, a master study investigated the brain structural connections between the cortical areas obtained from the fMRI study using diffusion MRI and the white matter query language.

7.2.5. White matter connectivity analysis in patients suffering from depression:

Participants: Julie Coloigner, Jean-Marie Batail, Jean-Christophe Ferré, Isabelle Corouge, Christian Barillot.

The mood depressive disorder (MDD) is a common chronically psychiatric disorder with an estimated lifetime prevalence reported to range from 10 percent to 15 percent worldwide. This disease is characterized by an intense dysregulation of affect and mood as well as additional abnormalities including cognitive dysfunction, insomnia, fatigue and appetite disturbance. Despite the extensive therapy options available for depression, up to 80 percent of patients will suffer from a relapse [1]. Consequently, exhibiting imaging biomarkers of this disease will support both a better understanding of the neural correlates underlying the depression, and a better diagnosis and treatment of individual depressed patients. Previous studies of structural and functional magnetic resonance imaging have reported several microstructural abnormalities in the prefrontal cortex, anterior cingulate cortex, hippocampus and thalamus [2]. These observations suggest a dysfunction of the circuits connecting frontal and subcortical brain regions, leading to a "disconnection syndrome" [3]. Given the small sample size used in the past studies, we proposed a more robust analysis using a larger cohort of patients suffering from depression, called LONGIDEP. The latter is a routine care cohort of patients suffering from mood depressive disorder who underwent a clinical evaluation, neuropsychological testing and brain MRI. The population sample consists of 125 patients suffering from depression and 65 healthy age and gender-matched, control subjects. A composite measure of medication load for each patient was assessed using a previously established method [4]. We investigated alterations of white matter integrity using a voxel-based analysis based on fractional anisotropy (FA) and the apparent diffusion coefficient (ADC) in patients with depression. Using graph theory-based analysis, we also examined white matter changes in the organization of networks in patients suffering from depression. Our findings provide robust evidence that the reduction of white-matter integrity in the interhemispheric connections and fronto-limbic neuronal circuits may play an important role in MDD pathogenesis. These results are consistent with an overall hypothesis that depression involves a disconnection of prefrontal, striatal, and limbic emotional areas.

7.2.6. Knowing and Remembering: Cognitive and Neural Influences of Familiarity on Recognition Memory in Early Alzheimer's Disease (EPMR-MA):

Participants: Pierre-Yves Jonin, Quentin Duché, Élise Bannier, Christian Barillot.

Inclusion of the 20 healthy participants in the "EPMR-MA" study (clinical trials ID NCT02492529) has been achieved, the inclusion phase will be achieved before 30th, december, 2017. Healthy controls data are pre-processed and the first analysis workflow proved promising, it should allow submitting a first paper at the beginning of 2018.

7.2.7. Semantic Dementia Imaging:

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

After demonstrating the relative preservation of fruit and vegetable knowledge in patients with semantic dementia (SD), we sought to identify the neural substrate of this unusual category effect. Nineteen patients with SD performed a semantic sorting task and underwent a morphometric 3T MRI scan. The grey-matter volumes of five regions within the temporal lobe were bilaterally computed, as well as those of two recently described areas (FG1 and FG2) within the posterior fusiform gyrus. In contrast to the other semantic categories we tested, fruit and vegetable scores were only predicted by left FG1 volume. We therefore found a specific relationship between the volume of a subregion within the left posterior fusiform gyrus and performance on fruits and vegetables in SD. We argue that the left FG1 is a convergence zone for the features that might be critical to successfully sort fruits and vegetables. We also discuss evidence for a functional specialization of the fusiform gyrus along two axes (lateral medial and longitudinal), depending on the nature of the concepts and on the level of processing complexity required by the ongoing task [28].

7.3. Research axis 3: Management of Information in Neuroimaging

Participants: Élise Bannier, Christian Barillot, Yao Chi, Isabelle Corouge, Olivier Commowick, Inès Fakhfakh, Michael Kain, Florent Leray, Julien Louis, Aneta Morawin, Mathieu Simon, Arnaud Touboulic.

The major topic that has been reached in the period concerns the sharing of data and processing tools in neuroimaging (through the ANR Neurolog and VIP projects, and more recently the "Programme d'Investissement d'Avenir" project such as OFSEP and FLI-IAM) that led to build a suitable architecture to share images and processing tools). Our overall goal within these projects was to set up a computational infrastructure to facilitate the sharing of neuroimaging data, as well as image processing tools, in a distributed and heterogeneous environment. These consortiums gathered expertises coming from several complementary domains: image processing in neuroimaging, workflows and grid computing, ontology development and ontology-based mediation. Shanoir (SHAring NeurOImaging Resources) is one of the major outcome of these projects. Shanoir uses semantics for concepts organization that are defined by the OntoNeuroLOG ontology. OntoNeuroLOG reuses and extends the OntoNeuroBase ontology. Both were designed using the same methodological framework, based on the use of the foundational ontology DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering), and the use of a number of core ontologies, that provide generic, basic and minimal concepts and relations in specific domains such as Artefacts. Shanoir aims at establishing the conditions allowing, through the Internet, to share distributed information sources in neuroimaging, whether these sources are located in various centers of experimentation, clinical departments of neurology, or research centers in cognitive neurosciences or image processing. 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. 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 Management (IAM) node of the France Life Imaging national infrastructure (FLI-IAM). Our team VisAGeS 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 infrastructure software solutions.

ASAP Project-Team

6. New Results

6.1. Theory of Distributed Systems

6.1.1. Simulation of Partial Replication in Distributed Transactional Memory

Participant: François Taïani.

Distributed Transactional Memory (DTM) is a concurrency mechanism aimed at simplifying distributed programming by allowing operations to execute atomically, mirroring the well-known transaction model of relational databases. DTM can play a fundamental role in the coordination of participants in mobile distributed applications. Most DTM solutions follow a full replication scheme, in spite of recent studies showing that partial replication approaches can present gains in scalability by reducing the amount of data stored at each node. This work [33] investigates the role of replica location in DTMs. The goal is to understand the effect of latency on the DTM's system performance in face of judicious replica distribution, taking into consideration the locations where data is more frequently accessed.

This work was performed in collaboration with with Diogo Lima and Hugo Miranda from the University of Lisbon (Portugal).

6.1.2. Distributed Universal Constructions: a Guided Tour

Participant: Michel Raynal.

The notion of a universal construction is central in computing science: the wheel has not to be reinvented for each new problem. In the context of n-process asynchronous distributed systems, a universal construction is an algorithm that is able to build any object defined by a sequential specification despite the occurrence of up to (n - 1) process crash failures. Michel Raynal presented a guided tour of such universal constructions in the bulletin of the EATCS [22]. Its spirit is not to be a catalog of the numerous constructions proposed so far, but a (as simple as possible) presentation of the basic concepts and mechanisms that constitute the basis these constructions rest on.

6.1.3. Atomic Read/Write Memory in Signature-Free Byzantine Asynchronous Message-Passing Systems

Participant: Michel Raynal.

This work introduced a signature-free distributed algorithm which builds an atomic read/write shared memory on top of a fully connected peer-to-peer 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 [42], 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 Mostéfaoui, Matoula Petrolia and Claude Jard from the University of Nantes and was published in Theory of Computing Systems [19].

6.1.4. From wait-free to arbitrary concurrent solo executions in colorless distributed computing

Participant: Michel Raynal.

In an asynchronous distributed system where any number of processes may crash, a process may have to run solo, computing its local output without receiving any information from other processes. In the basic shared memory system where the processes communicate through atomic read/write registers, at most one process may run solo.

In this work we introduced a new family of d-solo models, where d-processes may concurrently run solo, $1 \leq d \leq n$ (the 1-solo model is the basic read/write model). We studied distributed colorless computations in the d-solo models, where process ids are not used, either in task specifications or during computation, and we characterized the colorless tasks that can be solved in each d-solo model. Colorless tasks include consensus, set agreement and many other previously studied tasks. This shows that colorless algorithms have limited computational power for solving tasks, only when d>1. When d=1, colorless algorithms can solve the same tasks as algorithms that may use ids. It is well-known that, while consensus is not wait-free solvable in a model where at most one process may run solo, ϵ-approximate agreement is solvable. In a d-solo model, the fundamental solvable task is (d,ϵ)-solo approximate agreement, a generalization of ϵ-approximate agreement. Indeed, (d,ϵ)-solo approximate agreement can be solved in the d-solo model, but not in the (d+1)-solo model.

This work was carried out in collaboration with Maurice Herlihy from Brown University, Sergio Rajsbaum from UNAM (Mexico), and Julien Stainer from EPFL, in the context of the LIDICo associate team. It was published in Theoretical Computer Science [18].

6.1.5. Early Decision and Stopping in Synchronous Consensus: A Predicate-Based Guided Tour

Participant: Michel Raynal.

Consensus is the most basic agreement problem encountered in fault-tolerant distributed computing: each process proposes a value and non-faulty processes must agree on the same value, which has to be one of the proposed values. While this problem is impossible to solve in asynchronous systems prone to process crash failures, it can be solved in synchronous (round-based) systems where all but one process might crash in any execution. It is well-known that (t + 1) rounds are necessary and sufficient in the worst case execution scenario for the processes to decide and stop executing, where t < n is a system parameter denoting the maximum number of allowed process crashes and n denotes the number of processes in the system. Early decision and stopping considers the case where f < t processes actually crash, f not being known by processes. It has been shown that the number of rounds that have to be executed in the worst case is then min(f + 2, t + 1). In this work we showed that this value is an upper bound attained only in worst execution scenarios. This work resulted from a collaboration with Armando Castaneda from UNAM, Yoram Moses from Technion, and Matthieu Roy from LAAS Toulouse, in the context of the LIDICo associate team. It was published at NETYS 2017 [29].

6.1.6. Long-Lived Tasks

Participant: Michel Raynal.

The predominant notion for specifying problems to study distributed computability are tasks. Notable examples of tasks are consensus, set agreement, renaming and commit-adopt. The theory of task solvability is well-developed using topology techniques and distributed simulations. However, concurrent computing problems are usually specified by objects. Tasks and objects differ in at least two ways. While a task is a one-shot problem, an object, such as a queue or a stack, typically can be invoked multiple times by each process. Also, a task, defined in terms of sets, specifies its responses when invoked by each set of processes concurrently, while an object, defined in terms of sequences, specifies the outputs the object may produce when it is accessed sequentially.

In this work we showed how the notion of tasks can be extended to model any object. A potential benefit of this result is the use of topology, and other distributed computability techniques to study long-lived objects. This work resulted from a collaboration with Armando Castaneda and Sergio Rasjbaum from UNAM in the context of the LIDICo associate team. It was published at NETYS 2017 [35].

6.1.7. Which Broadcast Abstraction Captures k-Set Agreement?

Participant: Michel Raynal.

It is well-known that consensus (one-set agreement) and total order broadcast are equivalent in asynchronous systems prone to process crash failures. Considering wait-free systems, we addressed and answered the following question: which is the communication abstraction that "captures" k-set agreement? To this end, we introduced a new broadcast communication abstraction, called k-BO-Broadcast, which restricts the disagreement on the local deliveries of the messages that have been broadcast (1-BO-Broadcast boils down to total order broadcast). Hence, in this context, k=1 is not a special number, but only the first integer in an increasing integer sequence. This establishes a new "correspondence" between distributed agreement problems and communication abstractions, which enriches our understanding of the relations linking fundamental issues of fault-tolerant distributed computing. This work was carried out in collaboration with Damien Imbs from the University of Marseille, Achour Mostéfaoui from the University of Nantes, and Matthieu Perrin from IMDEA (Spain). It was published at DISC 2017 [39].

6.1.8. Signature-free asynchronous Byzantine systems: from multivalued to binary consensus with t< n/3, O(n2) messages, and constant time.

Participant: Michel Raynal.

We introduced 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 $t < n/3t < n/3$ (and is consequently optimal from a resilience point of view), uses $O(n^2)$ messages, has a constant time complexity, and uses neither signatures nor additional computational power (such as random numbers, failure detectors, additional scheduling assumption, or additional synchrony assumption). 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 satisfying the following property: if the set of a non-faulty process is a singleton containing value v, the set of any non-faulty process contains v. 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 $t < n/3t < n/3$ Byzantine processes, and without using cryptography techniques. Moreover, this reduction algorithm uses a single instance of the underlying binary consensus, and tolerates message re-ordering by Byzantine processes. This work, done in collaboration with Achour Mostefaoui from LS2N (Nantes), appeared in Acta Informatica [20].

6.1.9. A distributed leader election algorithm in crash-recovery and omissive system

Participant: Michel Raynal.

We introduced a new distributed leader election algorithm for crash-recovery and omission environments. Contrary to previous works, our algorithm tolerates the occurrence of crash-recoveries and message omissions to any process during some finite but unknown time, after which a majority of processes in the system remains up and does not omit messages. This work, done in collaboration with Christian Fernández-Campusano, Mikel Larrea, and Roberto Cortiñas from UPV/EHU, Spain, appeared in Information Processing Letters 2017 [16].

6.1.10. Providing Collision-Free and Conflict-Free Communication in General Synchronous Broadcast/Receive Networks

Participants: Michel Raynal, François Taïani.

This work [26] considers the problem of communication in dense and large scale wireless networks composed of resource-limited nodes. In this kind of networks, a massive amount of data is becoming increasingly available, and consequently implementing protocols achieving error-free communication channels constitutes an important challenge. Indeed, in this kind of networks, the prevention of message conflicts and message collisions is a crucial issue. In terms of graph theory, solving this issue amounts to solve the distance-2 coloring

problem in an arbitrary graph. The work presents a distributed algorithm providing the processes with such a coloring. This algorithm is itself collision-free and conflict-free. It is particularly suited to wireless networks composed of nodes with communication or local memory constraints.

This work was performed in collaboration with Abdelmadjid Bouabdallah and Hicham Lakhlef from Université Technologique de Compiègne (France).

6.1.11. Randomized abortable mutual exclusion with constant amortized RMR complexity on the CC model.

Participant: George Giakkoupis.

In [30], we presented an abortable mutual exclusion algorithm for the cache-coherent (CC) model with atomic registers and CAS objects. The algorithm has constant expected amortized RMR complexity in the oblivious adversary model and is deterministically deadlock-free. This is the first abortable mutual exclusion algorithm that achieves $o(\log n/\log\log n)$ RMR complexity.

This work was done in collaboration with Philipp Woelfel (University of Calgary).

6.2. Network and Graph Algorithms

6.2.1. Tight bounds on vertex connectivity under 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 [15], we proved the analogous result for vertex connectivity, when either vertices or edges are sampled. We showed 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. If edges are sampled with probability $p = \Omega(\log(n)/k)$then the sampled subgraph has vertex connectivity $\Omega(kp)$, with high probability. Both bounds are existentially optimal.

This work was done in collaboration with Keren Censor-Hillel (Technion), Mohsen Ghaffari (MIT), Bernhard Haeupler (Carnegie Mellon University), and Fabian Kuhn (University of Freiburg).

6.2.2. Tight bounds for coalescing-branching random walks on regular graphs

Participant: George Giakkoupis.

A *coalescing-branching random walk (Cobra)* is a natural extension to the standard random walk on a graph. The process starts with one pebble at an arbitrary node. In each round of the process every pebble splits into k pebbles, which are sent to k random neighbors. At the end of the round all pebbles at the same node coalesce into a single pebble. The process is also similar to randomized rumor spreading, with each informed node pushing the rumor to k random neighbors each time it receives a copy of the rumor. Besides its mathematical interest, this process is relevant as an information dissemination primitive and a basic model for the spread of epidemics.

In [25] we studied the *cover time* of Cobra walks, which is the time until each node has seen at least one pebble. Our main result is a bound of $O(\phi^{-1}\log n)$ rounds with high probability on the cover time of a Cobra walk with $k = 2$ on any regular graph with n nodes and conductance ϕ. This bound improves upon all previous bounds in terms of graph expansion parameters. Moreover, we showed that for any connected regular graph the cover time is $O(n\log n)$ with high probability, independently of the expansion. Both bounds are asymptotically tight.

This work was done in collaboration with Petra Berenbrink (University of Hamburg), Peter Kling (University of Hamburg).

6.3. Scalable Systems

6.3.1. Agar: A Caching System for Erasure-Coded Data

Participants: Anne-Marie Kermarrec, François Taïani.

Erasure coding is an established data protection mechanism. It provides high resiliency with low storage overhead, which makes it very attractive to storage systems developers. Unfortunately, when used in a distributed setting, erasure coding hampers a storage system's performance, because it requires clients to contact several, possibly remote sites to retrieve their data. This has hindered the adoption of erasure coding in practice, limiting its use to cold, archival data. Recent research showed that it is feasible to use erasure coding for hot data as well, thus opening new perspectives for improving erasure-coded storage systems. In this work [32], we address the problem of minimizing access latency in erasure-coded storage. We propose Agar—a novel caching system tailored for erasure-coded content. Agar optimizes the contents of the cache based on live information regarding data popularity and access latency to different data storage sites. Our system adapts a dynamic programming algorithm to optimize the choice of data blocks that are cached, using an approach akin to " Knapsack " algorithms. We compare Agar to the classical Least Recently Used and Least Frequently Used cache eviction policies, while varying the amount of data cached between a data chunk and a whole replica of the object. We show that Agar can achieve 16% to 41% lower latency than systems that use classical caching policies.

This work was performed in collaboration with from Raluca Halalai and Pascal Felber from Université de Neuchâtel (Switzerland).

6.3.2. Filament: A Cohort Construction Service for Decentralized Collaborative Editing Platforms

Participants: Resmi Ariyattu Chandrasekharannair, François Taïani.

Distributed collaborative editors allow several remote users to contribute concurrently to the same document. Only a limited number of concurrent users can be supported by the currently deployed editors. A number of peer-to-peer solutions have therefore been proposed to remove this limitation and allow a large number of users to work collaboratively. These approaches however tend to assume that all users edit the same set of documents, which is unlikely to be the case if such systems should become widely used and ubiquitous. In this work [24] we discuss a novel cohort-construction approach that allow users editing the same documents to rapidly find each other. Our proposal utilises the semantic relations between peers to construct a set of self-organizing overlays to route search requests. The resulting protocol is efficient, scalable, and provides beneficial load-balancing properties over the involved peers. We evaluate our approach and compare it against a standard Chord based DHT approach. Our approach performs as well as a DHT based approach but provides better load balancing.

6.3.3. Scalable Anti-KNN: Decentralized Computation of k-Furthest-Neighbor Graphs with HyFN

Participants: Simon Bouget, David Bromberg, François Taïani.

The decentralized construction of k-Furthest-Neighbor graphs has been little studied, although such structures can play a very useful role, for instance in a number of distributed resource allocation problems. In this work [27] we define KFN graphs; we propose HyFN, a generic peer-to-peer KFN construction algorithm, and thoroughly evaluate its behavior on a number of logical networks of varying sizes. 1 Motivation k-Nearest-Neighbor (KNN) graphs have found usage in a number of domains, including machine learning, recommenders, and search. Some applications do not however require the k closest nodes, but the k most dissimilar nodes, what we term the k-Furthest-Neighbor (KFN) graph. Virtual Machines (VMs) placement —i.e. the (re-)assignment of workloads in virtualised IT environments— is a good example of where KFN can be applied. The problem consists in finding an assignment of VMs on physical machines (PMs) that minimises some cost function(s). The problem has been described as one of the most complex and important for the IT industry, with large potential savings. An important challenge is that a solution does not only consist

in packing VMs onto PMs — it also requires to limit the amount of interferences between VMs hosted on the same PM. Whatever technique is used (e.g. clustering), interference aware VM placement algorithms need to identify complementary workloads — i.e. workloads that are dissimilar enough that the interferences between them are minimised. This is why the application of KFN graphs would make a lot of sense: identifying quickly complementary workloads (using KFN) to help placement algorithms would decrease the risks of interferences. The construction of KNN graphs in decentralized systems has been widely studied in the past. However, existing approaches typically assume a form of "likely transitivity" of similarity between nodes: if A is close to B, and B to C, then A is likely to be close to C. Unfortunately this property no longer holds when constructing KFN graphs. As a result, these approaches are not working anymore when applied to this new problem.

This work was performed in collaboration with Anthony Ventresque from University College Dublin (Ireland).

6.3.4. Density and Mobility-driven Evaluation of Broadcast Algorithms for MANETs

Participants: Simon Bouget, David Bromberg, François Taïani.

Broadcast is a fundamental operation in Mobile Ad-Hoc Networks (MANETs). A large variety of broadcast algorithms have been proposed. They differ in the way message forwarding between nodes is controlled, and in the level of information about the topology that this control requires. Deployment scenarios for MANETs vary widely, in particular in terms of nodes density and mobility. The choice of an algorithm depends on its expected coverage and energy cost, which are both impacted by the deployment context. In this work, we are interested in the comprehensive comparison of the costs and effectiveness of broadcast algorithms for MANETs depending on target environmental conditions. We did an experimental study of five algorithms, representative of the main design alternatives. Our study reveals that the best algorithm for a given situation, such as a high density and a stable network, is not necessarily the most appropriate for a different situation such as a sparse and mobile network. We identify the algorithms characteristics that are correlated with these differences and discuss the pros and cons of each design.

This work was done in collaboration with Etienne Rivière (University of Neuchatel), Laurent Réveillère (University of Bordeaux) and appeared in ICDCS 2017

6.3.5. An Adaptive Peer-Sampling Protocol for Building Networks of Browsers

Participant: Davide Frey.

Peer-sampling protocols constitute a fundamental mechanism for a number of large-scale distributed applications. The recent introduction of WebRTC facilitated the deployment of decentralized applications over a network of browsers. However, deploying existing peer-sampling protocols on top of WebRTC raises issues about their lack of adaptiveness to sudden bursts of popularity over a network that does not manage addressing or routing. In this contribution, we introduced SPRAY, a novel random peer-sampling protocol that dynamically, quickly, and efficiently self-adapts to the network size. We evaluated SPRAY by means of simulations and real-world experiments. This demonstrated its flexibility and highlighted its efficiency improvements at the cost of small overhead. We embedded SPRAY in a real-time decentralized editor running in browsers and ran experiments involving up to 600 communicating web browsers. The results demonstrate that SPRAY significantly reduces the network traffic according to the number of participants and saves bandwidth.

This work was carried out in collaboration with Brice Nédelec, Julian Tanke, Pascal Molli, and Achour Mostéfaoui from the University of Nantes and will appear in the World Wide Web Journal [21].

6.3.6. Designing Overlay Networks for Decentralized Clouds

Participant: Marin Bertier.

Recent increase in demand for next-to-source data processing and low-latency applications has shifted attention from the traditional centralized cloud to more distributed models such as edge computing. In order to fully leverage these models it is necessary to decentralize not only the computing resources but also their management. While a decentralized cloud has various inherent advantages, it also introduces different challenges with respect to coordination and collaboration between resources. A large-scale system with multiple administrative entities requires an overlay network which enables data and service localization based only on a partial view of the network. Numerous existing overlay networks target different properties but they are built in a generic context, without taking into account the specific requirements of a decentralized cloud. In this work [34], done in collaboration with G. Tato et C. Tedeschi from the Myriads project team, we identified some of these requirements and introduced Koala, a novel overlay network designed specifically to meet them.

ASCOLA Project-Team

7. New Results

7.1. Cloud programming and management

7.1.1. Cloud infrastructures

Our contributions regarding cloud infrastructures can be divided into three main topics described below: contributions related to (i) geo-distributed clouds (e.g., Fog and Edge computing), (ii) the convergence of Cloud and HPC infrastructures and (iii) the simulation of virtualized infrastructures.

7.1.1.1. Geo-distributed Clouds

Many academic and industry experts are now advocating a shift from large-centralized Cloud Computing infrastructures to massively small-geo-distributed data centers at the edge of the network. This new paradigm of utility computing is often called Fog and Edge Computing. Advantages of this paradigm are, among others, data-locality that enhances security aspects and response times for latency-critical applications, new energetic options because of reduced size of data centers (e.g., renewable energies), single point of failure avoidance etc. Among the obstacles to the adoption of this model though is the development of a convenient and powerful IaaS system capable of managing a significant number of remote data-centers in a unified way, including monitoring and data management issues in a decentralized environment.

In 2017, we achieved three main contributions toward this challenge.

In [12], we investigate how a holistic monitoring service for a Fog/Edge infrastructure, hosting next generation digital services, should be designed. Although several solutions have been proposed in the past for the monitoring of clusters, grids and cloud systems, none of those is well appropriate to the specific Fog and Edge Computing context. The contributions of this study are: (i) the problem statement, (ii) a classification and a qualitative analysis of major existing solutions, and (iii) a preliminary discussion of the impact of deployment strategies on the monitoring service.

In [6], [39], [17], we present successive studies related to the design and development of a first-class object store service for Fog/Edge facilities. After a deep analysis of major existing solutions (Ceph, Cassandra ...), we designed a proposal that combines Scale-out Network Attached Storage systems (NAS) and IPFS, a BitTorrent-based object store spread throughout the Fog/Edge infrastructure. Without impacting the IPFS advantages particularly in terms of data mobility, the use of a Scale-out NAS on each site reduces the inter-site exchanges that are costly but mandatory for the metadata management in the original IPFS implementation. Several experiments conducted on Grid'5000 testbed are analysed and corroborate, first, the benefit of using an object store service spread at the Edge, and second, the importance of mitigating inter-site accesses. Ongoing activities are related to the management of meta data information in order to benefit from data movements.

Finally, in [26], we introduce the premises of a fog/edge resource management system by leveraging the OpenStack software, a leading IaaS manager in the industry. The novelty of the presented prototype is to operate such an Internet-scale IaaS platform in a fully decentralized manner, using P2P mechanisms to achieve high flexibility and avoid single points of failure. More precisely, we revised the OpenStack Nova service (i.e., virtual machine management and allocation) by leveraging a distributed key/value store instead of the centralized SQL backend. We present experiments that validate the correct behavior and gives performance trends of our prototype through an emulation of several data-centers using Grid'5000 testbed.

7.1.1.2. Cloud and HPC convergence

Geo-distribution of Cloud Infrastructures is not the only current trend of utility computing. Another important challenge is to reach the convergence of Cloud and HPC infrastructures, in other words on-demand HPC. Among challenges of this convergence is, for example, the enhancement of the use of light virtualization techniques on HPC systems, as well as the enhancement of mechanisms to be able to consolidate those VMs without deteriorating the performance of HPC applications, thus minimizing interferences between applications.

In [36], we present Eley, a burst buffer solution that helps to accelerate the performance of Big Data applications while guaranteeing the QoS of HPC applications. To achieve this goal, Eley embraces interference-aware prefetching technique that makes reading data input faster while introducing low interference for HPC applications. Specifically, we equip the prefetcher with five optimization actions including No Action, Full Delay, Partial Delay, Scale Up and Scale Down. It iteratively chooses the best action to optimize the prefetching while guaranteeing the pre-defined QoS requirement of HPC applications (i.e., the deadline constraint for the completion of each I/O phase). Evaluations using a wide range of Big Data and HPC applications show the effectiveness of Eley in reducing the execution time of Big Data applications (shorter map phase) while maintaining the QoS of HPC applications.

7.1.1.3. Virtualization simulation

Finally, it is important to be able to simulate the behavior of proposals for the future architectures. However, current models for virtualized resources are not accurate.

In [32], we present our latest results regarding virtualization abstractions and models for cloud simulation toolkits. Cloud simulators still do not provide accurate models for most Virtual Machine (VM) operations. This leads to incorrect results in evaluating real cloud systems. Following previous works on live-migration, we discuss an experimental study we conducted in order to propose a first-class VM boot time model. Most cloud simulators often ignore the VM boot time or give a naive model to represent it. After studying the relationship between the VM boot time and different system parameters such as CPU utilization, memory usage, I/O and network bandwidth, we introduce a first boot time model that could be integrated into current cloud simulators. Through experiments, we also show that our model correctly reproduced the boot time of a VM under different resources contention.

7.1.2. Deployment and reconfiguration in the Cloud

Being able to manage the new generation of utility computing infrastructures is an important step to build useful system building blocks. The next step is to be able to perform initial deployment of any kind of distributed software (i.e., systems, frameworks or applications) on those infrastructures, thus dealing with a complex process that includes interactions between building blocks such as virtual machine management, optimized deployment plans, monitoring of deployment etc. Such deployment processes cannot be handled manually anymore, for this reason automatic deployments tools have to be designed according to the challenges of new infrastructures (e.g., geo-distribution, hybrid infrastructures etc.). Moreover, as distributed software are more and more dynamic (i.e., reconfiguring themselves at runtime), reconfiguration and self-management capabilities should be handled in an efficient and scalable manner.

7.1.2.1. Initial deployment and placement strategies

When focusing on the initial deployment, many challenges should already need to be addressed such as placement of distributed software onto virtual machines, themselves being placed onto physical resources. This kind of placement problem can be modeled in many different ways, such as linear or constraint programming or graph partitioning. Most of the time a multi-objective NP-hard problem is formulated, and specific heuristics have to be built to reach scalable solutions.

In [18], we present new specific placement constraints and objectives adapted to hybrid clouds infrastructures, and we address this problem through constraint programming. Furthermore we evaluate the expressivity and performance of the solution on a real case study. In the Cloud, if public providers enable simple access to resources for companies and users who have sporadic computation or storage needs, private clouds could sometimes be preferred for security or privacy reasons, or for cost reasons due to a high frequency usage of services. However, in many cases a choice between public or private clouds does not fulfill all requirements of companies and hybrid cloud infrastructures should be preferred. Solutions have already been proposed to address hybrid cloud infrastructures, however most of the time the placement of a distributed software on such infrastructure has to be indicated manually.

In [37], we present a geo-aware graph partitioning method named G-Cut, which aims at minimizing the inter-DC data transfer time of graph processing jobs in geo-distributed DCs while satisfying the WAN usage budget. G-Cut adopts two novel optimization phases which address the two challenges in WAN usage and network heterogeneities separately. G-Cut can be also applied to partition dynamic graphs thanks to its light-weight runtime overhead. We evaluate the effectiveness and efficiency of G-Cut using real-world graphs with both real geo-distributed DCs and simulations. Evaluation results demonstrate that effectiveness of G-Cut in reducing the inter-DC data transfer time and the WAN usage with a low runtime overhead.

Many other challenges than placement rise from the initial deployment. In [20], we present a survey of existing deployment tools that have been used in production to deploy OpenStack, which is a complex distributed system composed of more than a hundred different services. To fully understand how IaaSes are deployed today, we propose in this paper an overall model of the application deployment process that describes each step with their interactions. This model then serves as the basis to analyse five different deployment tools used to deploy OpenStack in production: Kolla, Enos, Juju, Kubernetes, and TripleO. Finally, a comparison is provided and the results are discussed to extend this analysis.

7.1.2.2. Capacity planning and scheduling

While a placement problem is a discrete problem at a given instant, some other challenges of deployment and reconfiguration may include the time dimension leading to scheduling optimization.

in [30] we have proposed two original workload prediction models for Cloud infrastructures. These two models, respectively based on constraint programming and neural networks, focus on predicting the CPU usage of physical servers in a Cloud data center. The predictions could then be exploited for designing energy-efficient resource allocation mechanisms like scheduling heuristics or over-commitment policies. We also provide an efficient trace generator based on constraint satisfaction problem and using a small amount of real traces. Such a generator can overcome availability issues of extensive real workload traces employed for optimization heuristics validation. While neural networks exhibit higher prediction capabilities, constraint programming techniques are more suitable for trace generation, thus making both techniques complementary.

7.1.2.3. Reconfiguration and self-management

Being able to handle the dynamicity of hardware, system building blocks, middleware and applications is a great challenge of today's and future utility computing systems. On large infrastructures such as Cloud, Fog or Edge Computing, manual administration of such dynamicity is not feasible. The automatic management of reconfiguration, or self-management of software is of great importance to guarantee reliability, fault tolerance, security, and cost and energy optimization.

In [4], in order to improve the self-adaptive behaviors in the context of Component-based Architecture, we design self-adaptive software components based on logical discrete control approaches, in which the self-adaptive behavioural models enrich component controllers with a knowledge not only on events, configurations and past history, but also with possible future configurations. This article provides the description, implementation and discussion of Ctrl-F, a Domain-specific Language whose objective is to provide high-level support for describing these control policies. In [13], we extended Ctrl-F with modularity capabilities. Apart from the benefits of reuse and substitutability of Ctrl-F programs, modularity allows to break down the combinatorial explosion intrinsic to the generation of correct-by-construction controllers in the compilation process of Ctrl-F. A further advantage of modularity is that the executable code, that is, the controllers resulting from that compilation, are loss-coupled and can therefore be deployed and executed in a distributed fashion.

However, higher abstraction-level tools also have to be proposed for reconfiguration. In [21], we introduce ElaScript, a Domain Specific Language (DSL) which offers Cloud administrators a simple and concise way to define complex elasticity-based reconfiguration plans. ElaScript is capable of dealing with both infrastructure and software elasticities, independently or together, in a coordinated way. We validate our approach by first showing the interest to have a DSL offering multiple levels of control for Cloud elasticity, and then by showing its integration with a realistic well-known application benchmark deployed in OpenStack and the Grid'5000 infrastructure testbed.

Finally, self-management can be applied at many different levels of the Cloud paradigm, from infrastructure reconfigurations to application topology reconfigurations. In practice these reconfiguration mechanisms are tightly coupled. For example, a change in the infrastructure could lead to the re-deployment of virtual machines upon it that could lead itself to application reconfigurations. In [27], we advocate that Cloud services, regardless of the layer, may share the same consumer/provider-based abstract model. From that model, we can derive a unique and generic Autonomic Manager (AM) that can be used to manage any XaaS (Everything-as-a-Service) layer defined with that model. The paper proposes such an abstract (although extensible) model along with a generic constraint-based AM that reasons on abstract concepts, service dependencies as well as SLA (Service Level Agreements) constraints in order to find the optimal configuration for the modeled XaaS. The genericity of our approach are shown and discussed through two motivating examples and a qualitative experiment has been carried out in order to show the applicability of our approach as well as to discuss its limitations.

7.2. Energy-aware computing

7.2.1. Renewable energy

In his PhD thesis [1], Md Sabbir Hasan proposes – across three different contributions – how to smartly use green energy at the infrastructure and application levels for further reduction of the corresponding carbon footprints. First, he investigates the options and challenges to integrate different renewable energy sources in a realistic way and proposes a *Cloud energy broker*, which can adjust the availability and price combination to buy Green energy dynamically from the energy market in advance to make a data center partially green. Then, he introduces the concept of *virtualization of green energy*, which can be seen as an alternative to energy storage used in data centers to eliminate the intermittency problem to some extent. With the adoption of this virtualization concept, we can maximize the usage of green energy contrary to energy storage which induces energy losses, while introducing a notion of Green Service Level Agreement based on green energy for service provider and end-users. Finally, he proposes an energy adaptive autoscaling solution to exploit application internals to create green energy awareness in the interactive SaaS applications, while respecting traditional QoS properties.

In [9], we present a scheme for green energy management in the presence of explicit and implicit integration of renewable energy in data center. 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 [0] to support Green SLA by introducing two new threshold parameters and iii) we introduce green SLA 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.

In [8], we investigate a thorough analysis of energy consumption and performance trade-off by allowing smart usage of green energy for interactive cloud application. Moreover, we propose an auto-scaler, named as SaaScaler, that implements several control loop based application controllers to satisfy different performance (i.e., response time, availability and user experience) and resource aware metrics (i.e., quality of energy). Based on extensive experiments with RUBiS benchmark and real workload traces using single compute node in Openstack/Grid'5000, results suggest that 13% brown energy consumption can be reduced without deprovisioning any physical or virtual resources at IaaS layer while 29% more users can access the application by dynamically adjusting capacity requirements. In [23], we add to the previous paper the capability of the infrastructure layer to be elastic. We propose a PaaS solution which efficiently utilize the elasticity nature at both infrastructure and application levels, by leveraging adaptation in facing to changing condition i.e., workload burst, performance degradation, quality of energy, etc. While applications are adapted by dynamically re-configuring their service level based on performance and/or green energy availability, the infrastructure takes care of addition/removal of resources based on application's resource demand. Both

[0] http://web.imt-atlantique.fr/x-info/csla

adaptive behaviors are implemented in separated modules and are coordinated in a sequential manner. We validate our approach by extensive experiments and results obtained over Grid'5000 testbed. Results show that, application can reduce significant amount of brown energy consumption by 35% and daily instance hour cost by 37% compared to a baseline approach.

in [28] we address the problem of improving the utilization of renewable energy for a single data center by using two approaches: opportunistic scheduling and energy storage. Our first result deals with analyzing the workload to find ideal solar panel dimension and battery size, this is used to power the entire workload without any brown energy consumption. However, in reality, either the solar panel dimension or the battery size are limited, and we still have to address the problem of matching the workload consumption and renewable energy production. The second result shows that opportunistic scheduling can reduce the demand for battery size while the renewable energy is sufficient. The last results demonstrate that for different battery sizes and solar panel dimensions, we can find an optimal solution combining both approaches that balances the energy losses due to different causes such as battery efficiency and VM migrations due to consolidation algorithms.

In [5] we presented the EPOC project, focus on energy-aware task execution from the hardware to application's components in the context of a mono-site data center (all resources are in the same physical location) which is connected to the regular electric Grid and to renewable energy sources (such as windmills or solar cells). we have presented the EpoCloud principles, architecture and middleware components. EpoCloud is our prototype, which tackles three major challenges: 1) To optimize the energy consumption of distributed infrastructures and service compositions in the presence of ever more dynamic service applications and ever more stringent availability requirements for services; 2) To design a clever cloud's resource management, which takes advantage of renewable energy availability to perform opportunistic tasks, then exploring the trade-off between energy saving and performance aspects in large-scale distributed system; 3) To investigate energy-aware optical ultra high-speed interconnection networks to exchange large volumes of data (VM memory and storage) over very short periods of time.

in [31] we extend our previous work on PIKA (focus 2 in the EPOC project) and introduced the green energy aware scheduling problem (GEASP) to optimize the energy consumption of a small/medium size data center. Using our model to solve the GEASP, we could optimize the energy consumption of a small/medium size data center in three ways. First, we slightly decrease its overall energy consumption, second we considerably decrease its brown energy consumption and finally we significantly increase its green energy consumption.

7.2.2. Energy-aware consolidation and reconfiguration

In [41] we compared the performance of VMs and containers when consolidating multiple services, in terms of QoS and EE. Our experiments compared two broadly recognized virtualization technologies: KVM for the VM approach, and Docker for the containers. We conclude that Docker outperforms KVM both in QoS and EE. According to our measurements, Docker allows running up to a 21% more services than KVM, when setting a maximum latency of 3,000 ms. In this configuration, Docker offers this service while using a 11.33% less energy than KVM. At a datacenter level, the same computation could run using less servers and less energy per server, accounting for a total of a 28% energy savings inside the datacenter.

The emergence of Internet of Things (IoT) is participating to the increase of data- and energy-hungry applications. As connected devices do not yet offer enough capabilities for sustaining these applications, users perform computation offloading to the cloud. To avoid network bottlenecks and reduce the costs associated to data movement, edge cloud solutions have started being deployed, thus improving the Quality of Service. In [29] , we advocated for leveraging on-site renewable energy production in the different edge cloud nodes to green IoT systems while offering improved QoS compared to core cloud solutions. We proposed an analytic model to decide whether to offload computation from the objects to the edge or to the core Cloud, depending on the renewable energy availability and the desired application QoS. This model is validated on our application use-case that deals with video stream analysis from vehicle cameras.

In [33], we address the problem of stragglers (i.e., slow tasks) in Big Data applications. In particular, we introduce a novel straggler detection mechanism to improve the energy efficiency of speculative execution in Hadoop, namely a hierarchical detection mechanism. The goal of this detection mechanism is to identify

critical stragglers which strongly affect the job execution times and reduce the number of killed speculative copies which lead to energy waste. We also present an energy-aware copy allocation method to reduce the energy consumption of speculative execution. The core of this allocation method is a performance model and an energy model which expose the trade-off between performance and energy consumption when scheduling a copy. We evaluate our hierarchical detection mechanism and energy-aware copy allocation method on the Grid'5000 testbed using three representative MapReduce applications. Experimental results show a good reduction in the resource wasted on killed speculative copies and an improvement in the energy efficiency compared to state-of-the-art mechanisms.

The increasing size of main memories has lead to the advent of new types of storage systems. These systems propose to keep all data in distributed main memories. In [35], we present a study to characterize the performance and energy consumption of a representative in-memory storage system, namely RAMCloud, to reveal the main factors contributing to performance degradation and energy-inefficiency. Firstly, we reveal that although RAMCloud scales linearly in throughput for read-only applications, it has a non-proportional power consumption. Mainly because it exhibits the same CPU usage under different levels of access. Secondly, we show that prevalent Web workloads i.e., read-heavy and update-heavy workloads, can impact significantly the performance and the energy consumption. We relate it to the impact of concurrency, i.e., RAMCloud poorly handles its threads under highly-concurrent accesses. Thirdly, we show that replication can be a major bottleneck for performance and energy. Finally, we quantify the overhead of the crash-recovery mechanism in RAMCloud on both energy-consumption and performance.

7.3. Software engineering

7.3.1. Security and privacy

This year, we have developed new results on the security and privacy of cloud systems on all layers of abstraction: a first notion of distributed side-channel attacks on the system-level, privacy-aware middleware storage systems and accountability specifications and implementations on the application level.

7.3.1.1. System-level security for virtualized environments

Isolation on the system-level is a core security challenge for Cloud infrastructures. Similarly, fog and edge infrastructures are based on virtualization to share physical resources among several self-contained execution environments like virtual machines and containers. Yet, isolation may be threatened due to side-channels, created by the virtualization layer or due to the sharing of physical resources like the processor. Side-channel attacks (SCAs) exploit and use such leaky channels to obtain sensitive data. Previous SCAs are local and exploit isolation challenges of virtualized environments to retrieve sensitive information. We have introduced, as a first, the concept of *distributed side-channel attack (DSCA)* that is based on coordinating local attack techniques. We have explored how such attacks can threaten isolation of any virtualized environments such as fog and edge computing. Finally, we have proposed a first set of applicable countermeasures for attack mitigation of DSCAs. [14], [44]

In [24] we presented how the increasing adoption of cloud environments operated with virtualization technology opened the way to a promising hypervisor-based security monitoring approach named Virtual Machine Introspection (VMI). We investigated in Kbin-ID the application of binary code introspection at hypervisor level and analysis mechanisms on all VM kernel binary code, namely all kernel functions, to widely narrow the semantic gap in an automatic and largely OS independent way. Kbin-ID [40] is a novel hypervisor-based main kernel binary code disassembler which enables the hypervisor to locate all VM main kernel binary code and divide it into code blocks given only the address of one arbitrary kernel instruction. In [24] we presented a security use case, we are able to detect running processes that are hidden from Linux task list and ps command output, and more generally that our solution can be used for designing easily automatic and largely kernel portable VMI applications that detect and safely react against malicious activities thanks to the instrumentation of kernel functions.

7.3.1.2. Privacy-Aware Data Storage.

In [34] we propose a cloud storage service that protects the privacy of users by breaking user documents into blocks in order to spread them on several cloud providers. As cloud providers only own a part of the blocks and they do not know the block organization, they can not read user documents. Moreover, the storage service connects directly users and cloud providers without using a third-party as is generally the practice in cloud storage services. Consequently, users do not give critical information (security keys, passwords, etc.) to a third-party.

7.3.1.3. Accountability for Cloud applications.

Nowadays we are witnessing the democratization of cloud services, as a result, more and more end-users (individuals and businesses) are using these services in their daily life. In such scenarios, personal data is generally flowed between several entities. end-users need to be aware of the management, processing, storage and retention of personal data, and to have necessary means to hold service providers accountable for the use of their data. In Walid Benghabrit's thesis we present an accountability framework called Accountability Laboratory (AccLab) that allows to consider accountability from design time to implementation. We developed a language called Abstract Accountability Language (AAL) that allows to write obligations and accountability policies. This language is based on a formal logic called First Order Linear Temporal Logic (FOTL) which allows to check the consistency of the accountability policies and the compliance between two policies. These policies are translated into a temporal logic called FO-DTL 3, which is associated to a monitoring technique based on formula rewriting. Finally we developed a monitoring tool called Accountability Monitoring (AccMon) which provides means to monitor accountability policies in the context of a real system. These policies are based on FO-DTL 3 logic and the framework can act in both centralized and distributed modes and can run in on-line and off-line modes.

Accountability means to obey a contract and to ensure responsibilities in case of violations. In previous work we defined the Abstract Accountability Language and its AccLab tool support. In order to evaluate the suitability of our language and tool we experiment with the laptop user agreement, one of the policies of the Hope University in Liverpool. While this experiment is still incomplete we are able to draw some preliminary conclusions. The use of FOTL is rather tricky and the only existing prover is not maintained we think to target a first-order logic approach in the future. Natural specifications have traditional issues, for instance missing information, noises, ambiguities etc. But in case of these policies we can say much more. The information system is missing but also most of the details about the auditing process and the rectification aspects (sanction, compensation, explanation, etc). There is also a mixture of proper user behavior with the usage policy which confuses the specifier. A mean to structure the specification is important, we suggest to use templates, and it is also convenient to capture usage and accountability practices.

7.3.2. Software development and programming languages

7.3.2.1. Industrial Internet

In [19], we present a first "vision" paper toward Cloud Manufacturing. More precisely we try to reconsider relationships between Cloud Computing and Cloud Manufacturing based on basic definitions and historical evolution of both worlds. History shows many relations between computer science and manufacturing processes, starting with the initial idea of "digital manufacturing" in the '70s. Since then, advances in computer science have given birth to the *Cloud Computing* (CC) paradigm, where computing resources are seen as a *service* offered to various end-users. Of course, CC has been used as such to improve the IT infrastructure associated to a manufacturing infrastructure, but its principles have also inspired a new manufacturing paradigm *Cloud Manufacturing* (CMfg) with the perspective of many benefits for both the manufacturers and their customers. However, despite the usefulness of CC for CMfg, we advocate that considering CC as a core enabling technology for CMfg, as is often put forth in the literature, is limited and should be reconsidered. This paper presents a new core-enabling vision toward CMfg, called *Cloud Anything* (CA). CA is based on the idea of abstracting low-level resources, beyond computing resources, into a set of core control building blocks providing the grounds on top of which any domain could be "cloudified".

7.3.2.2. Cloud and HPC programming

In [43], we deal with testing reproducibility in the context of Cloud elasticity, which requires control of the elasticity behavior, the possibility to select specific resources to be allocated/unallocated, and the coordination of events parallel to the elasticity process. We propose an approach fulfilling those requirements in order to make elasticity testing reproducible. To validate our approach, we perform three experiments on representative bugs on MongoDB and Zookeeper Cloud applications, where our approach succeeds in reproducing all the bugs.

In [7], the Multi-Stencil Framework (MSF) is presented. Even though this framework is applied on HPC numerical simulations, this work can be transposed to many different domains, for instance smart-* applications of Fog and Edge computing infrastructures, where heterogeneity of computations and programming models have to be handled. As the computation power of modern high performance architectures increases, their heterogeneity and complexity also become more important. One of the big challenges of exascale is to reach programming models that give access to high performance computing (HPC) to many scientists and not only to a few HPC specialists. One relevant solution to ease parallel programming for scientists is Domain Specific Language (DSL). However, one problem to avoid with DSLs is to mutualized existing codes and libraries instead of implementing each solution from scratch. For example, this phenomenon occurs for stencil-based numerical simulations, for which a large number of languages has been proposed without code reuse between them. The Multi-Stencil Framework (MSF) presented in this paper combines a new DSL to component-based programming models to enhance code reuse and separation of concerns in the specific case of stencils. MSF can easily choose one parallelization technique or another, one optimization or another, as well as one back-end implementation or another. It is shown that MSF can reach same performances than a non component-based MPI implementation over 16.384 cores. Finally, the performance model of the framework for hybrid parallelization is validated by evaluations.

DIONYSOS Project-Team

7. New Results

7.1. Performance Evaluation

Participants: Yann Busnel, Yves Mocquard, Bruno Sericola, Gerardo Rubino

Correlation estimation between distributed massive streams. The real time analysis of massive data streams is of utmost importance in data intensive applications that need to detect as fast as possible and as efficiently as possible (in terms of computation and memory space) any correlation between its inputs or any deviance from some expected nominal behavior. The IoT infrastructure can be used for monitoring any events or changes in structural conditions that can compromise safety and increase risk. It is thus a recurrent and crucial issue to determine whether huge data streams, received at monitored devices, are correlated or not as it may reveal the presence of attacks. In [14] we propose a metric, called *Codeviation*, that allows to evaluate the correlation between distributed massive streams. This metric is inspired from classical material in statistics and probability theory, and as such enables 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 domain from which data items are drawn. We then generalize our approach by presenting a new metric, the *Sketch-☆ metric*, which allows us to define a distance between updatable summaries of large data streams. An important feature of the *Sketch-☆ metric* is that, given a measure on the entire initial data streams, the *Sketch-☆ metric* preserves the axioms of the latter measure on the sketch. We also conducted extensive experiments on both synthetic traces and real data sets allowing us to validate the robustness and accuracy of our metrics.

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.

Shuffle grouping is a technique used by stream processing frameworks to share input load among parallel instances of stateless operators. With shuffle grouping each tuple of a stream can be assigned to any available operator instance, independently from any previous assignment. A common approach to implement shuffle grouping is to adopt a Round-Robin policy, a simple solution that fares well as long as the tuple execution time is almost the same for all the tuples. However, such an assumption rarely holds in real cases where execution time strongly depends on tuple content. As a consequence, parallel stateless operators within stream processing applications may experience unpredictable unbalance that, in the end, causes undesirable increase in tuple completion times. In [61] we propose Online Shuffle Grouping (OSG), a novel approach to shuffle grouping aimed at reducing the overall tuple completion time. OSG estimates the execution time of each tuple, enabling a proactive and online scheduling of input load to the target operator instances. Sketches are used to efficiently store the otherwise large amount of information required to schedule incoming load. We provide a probabilistic analysis and illustrate, through both simulations and a running prototype, its impact on stream processing applications.

Grand Challenge. Since 2011, the ACM International Conference on Distributed Event-based Systems (DEBS) launched the Grand Challenge series to increase the focus on these systems as well as provide common benchmarks to evaluate and compare them. The ACM DEBS 2017 Grand Challenge focused on (soft) real-time anomaly detection in manufacturing equipment. To handle continuous monitoring, each machine is fitted with a vast array of sensors, either digital or analog. These sensors provide periodic measurements, which are sent to a monitoring base station. The latter receives then a large collection of observations. Analyzing in an efficient and accurate way, this very-high-rate – and potentially massive – stream of events is the core of the Grand Challenge. Although, the analysis of a massive amount of sensor reading requires an on-line analytics pipeline that deals with linked-data, clustering as well as a Markov model training and querying. The FlinkMan system [62] proposes a solution to the 2017 Grand Challenge, making use of a publicly available streaming engine and thus offering a generic solution that is not specially tailored for this or for another challenge. We offer an efficient solution that maximally utilizes available cores, balances the load among the cores, and avoids to the extent possible tasks such as garbage collection that are only indirectly related to the task at hand.

Health big data processing. Sharing and exploiting efficiently Health Big Data (HBD) lead to tackle great challenges: data protection and governance taking into account legal, ethical and deontological aspects which enables a trust, transparent and win-to-win relationship between researchers, citizen and data providers. Lack of interoperability: data are compartmentalized and are so syntactically and semantically heterogeneous. Variable data quality with a great impact on data management and statistical analysis. The objective of the INSHARE project [41] is to explore, through an experimental proof of concept, how recent technologies could overcome such issues. It aims at demonstrating the feasibility and the added value of an IT platform based on CDW, dedicated to collaborative HBD sharing for medical research.

The consortium includes 6 data providers: 2 academic hospitals, the SNIIRAM (the French national reimbursement database) and 3 national or regional registries. The platform is designed following a three steps approach: (1) to analyze use cases, needs and requirements, (2) to define data sharing governance and secure access to the platform, (3) to define the platform specifications. Three use cases (healthcare trajectory analysis, epidemiological registry enrichment, signal detection) were analyzed to design the platform corresponding to five studies and using eleven data sources. The governance was derived from the SCANNER model and adapted to data sharing. As a result, the platform architecture integrates the following tools and services: data repository and hosting, semantic integration services, data processing, aggregate computing, data quality and integrity monitoring, id linking, multi-source query builder, visualization and data export services, data governance, study management service and security including data watermarking.

Throughput prediction in cellular networks. Downlink data rates can vary significantly in cellular networks, with a potentially non-negligible effect on the user experience. Content providers address this problem by using different representations (*e.g.*, picture resolution, video resolution and rate) of the same content and by switching among these based on measurements collected during the connection. If it were possible to know the achievable data rate before the connection establishment, content providers could choose the most appropriate representation from the very beginning. We have conducted a measurement campaign involving 60 users connected to a production network in France, to determine whether it is possible to predict the achievable data rate using measurements collected, before establishing the connection to the content provider, on the operator's network and on the mobile node. We show that it is indeed possible to exploit these measurements to predict, with a reasonable accuracy, the achievable data rate [53].

Population protocol model. We consider in [50] a large system populated by n anonymous nodes that communicate through asynchronous and pairwise interactions. The aim of these interactions is, for each node, to converge toward a global property of the system that depends on the initial state of the nodes. We focus on both the counting and proportion problems. We show that for any $\delta \in (0,1)$, the number of interactions needed per node to converge is $O(\ln(n/\delta))$ with probability at least $1-\delta$. We also prove that each node can determine, with any high probability, the proportion of nodes that initially started in a given state without knowing the number of nodes in the system. This work provides a precise analysis of the convergence bounds, and shows that using the 4-norm is very effective to derive useful bounds.

The context of [71] is the well studied dissemination of information in large scale distributed networks through pairwise interactions. This problem, originally called *rumor mongering*, and then *rumor spreading* has mainly been investigated in the synchronous model, which relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In this paper, we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where, at random times, nodes successively interact by pairs exchanging their information on the rumor. In a previous paper, we performed a study of the total number of interactions needed for all the nodes of the network to discover the rumor. While most of the existing results involve huge constants that do not allow us to compare different protocols, we provided a thorough analysis of the distribution of this total number of interactions together with its asymptotic behavior. In this paper we extend this discrete-time analysis by solving a conjecture proposed previously and we consider the continuous-time case, where a Poisson process is associated with each node to determine the instants at which interactions occur. The rumor spreading time is thus more realistic since it is the time needed for all the nodes of the network to discover the rumor. Once again, as most of the existing results involve huge constants, we provide a tight bound and equivalent of the complementary distribution of the rumor spreading time. We also give the exact asymptotic behavior of the complementary distribution of the rumor spreading time around its expected value when the number of nodes tends to infinity.

Transient analysis. Last, in two keynotes ([35] and [34]), we described part of our previous analytical results concerning the transient behavior of well-structured Markov processes, mainly on performance models (queueing systems), and we presented recent new results that extend those initial findings. The heart of the novelties lie on an extension of the concept of duality proposed by Anderson in [73] that we call pseudo-dual. The dual of a stochastic process needs strong monotonicity conditions to exist. Our proposed pseudo-dual always exist, and is directly defined on a linear system of differential equations with constant coefficients, that can be, in particular, the system of Chapman-Kolmogorov equations corresponding to a Markov process, but not necessarily. This allows, for instance, to prove the validity of closed-forms expressions of the transient distribution of a Markov process in cases where the dual doesn't exist. The keynote [35] was presented to a public oriented toward differential equations and dynamical systems; [34] has a more modeling flavour. A paper is under preparation with the technical details.

7.2. Distributed deep learning on edge-devices

Participants: Corentin Hardy, Gerardo Rubino, Bruno Sericola

A large portion of data mining and analytic services use modern machine learning techniques, such as deep learning. The state-of-the-art results related to deep learning come at the price of an intensive use of computing resources. The leading frameworks (*e.g.*, TensorFlow) are executed on GPUs or on high-end servers in data centers. On the other end, there is a proliferation of personal devices with possibly free CPU cycles; this can enable services to run in users' homes, embedding machine learning operations. In [66] and [43], we ask the following question: *Is distributed deep learning computation on WAN connected devices feasible, in spite of the traffic caused by learning tasks*? We show that such a setup rises some important challenges, most notably the ingress traffic that the servers hosting the up-to-date model have to sustain. In order to reduce this stress, we propose *AdaComp*, a novel algorithm for compressing worker updates to the model on the server. Applicable to stochastic gradient descent based approaches, it combines efficient gradient selection and learning rate modulation. We experiment and measure the impact of compression, device heterogeneity and reliability on the accuracy of learned models, with an emulator platform that embeds TensorFlow into Linux containers. We report a reduction of the total amount of data sent by workers to the server by two order of magnitude (*e.g.*, 191-fold reduction for a convolutional network on the MNIST dataset), when compared to a standard asynchronous stochastic gradient descent, while preserving model accuracy. The extension of the AdaComp algorithm to Random Neural Networks started with the introduction of Random Neural Layers, see [65].

7.3. Network Economics

Participants: Bruno Tuffin, Patrick Maillé, Pierre L'Ecuyer

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 [7].

Roaming. In October 2015, the European parliament has decided to forbid roaming charges among EU mobile phone users, starting June 2017, as a first step toward the unification of the European digital market. We have investigated the consequences of such a measure from an economic perspective. In [47], we analyze the effect of the willingness-to-pay heterogeneity among users (also due to wealth heterogeneity), and the fact that the roaming behavior is positively correlated with wealth. Our analysis suggests that imposing free roaming degrades the revenues of the operator but can also deter some users from subscribing; hence we conclude that such (apparently beneficial) regulatory decisions must be taken with care. In [47], we particularly focus on the strategies on transit payments between ISPs in different countries. We highlight that scrutiny is also required since, depending on parameters, consumer surplus or subscription penetration are not necessarily maximized if free roaming is enforced.

Network neutrality. Most of our activity has been devoted to the vivid network neutrality debate, going beyond the traditional for or against neutrality, and trying to tackle it from different angles.

Network neutrality has been a very sensitive topic of discussion all over the world. In the keynote talk [59], we first introduce the elements of the debate and how the problem can be modeled and analyzed through game theory. With an Internet ecosystem much more complex now than the simple delivery chain Content-ISP-User, we highlight, in a second step, how neutrality principles can be bypassed in various ways without violating the rules currently evoked in the debate, for example via Content Delivery Networks (CDNs), or via search engines which can affect the visibility and accessibility of content. We describe some other grey zones requiring to be dealt with and spend some time on discussing the (potential) implications for clouds.

The impact of CDNs on the debate has been detailed in [18]. Content Delivery Networks (CDN) have become key telecommunication actors. They contribute to improve significantly the quality of services delivering content to end users. However, their impact on the ecosystem raises concerns about their fairness, and therefore the question of their inclusion in the neutrality debates becomes relevant. We analyze the impact of a revenue-maximizing CDN on some other major actors, namely, the end-users, the network operators and the content providers, at comparing the outcome with that of a fair behavior, and at providing tools to investigate whether some regulation should be introduced.We present a mathematical model and show that there exists a unique optimal revenue-maximizing policy for a CDN actor, in terms of dimensioning and allocation of its storage capacity, and depending on parameters such as prices for service/transport/storage. Numerical experiments are then performed with both synthetic data and real traces obtained from a major Video-on-Demand provider. In addition, using the real traces, we compare the revenue-based policy with policies based on several fairness criteria.

Network neutrality is often advocated by content providers, stressing that side payments to Internet Service Providers would hinder innovation. However, we also observe some content providers actually paying those fees. In [24], we intend to explain such behaviors through economic modeling, illustrating how side payments can be a way for an incumbent content provider to prevent new competitors from entering the market. We investigate the conditions under which the incumbent can benefit from such a barrier-to-entry, and the consequences of that strategic behavior on the other actors: content providers, users, and the Internet Service Provider. We also describe how the Nash bargaining solution concept can be used to determine the side payment.

Similarly, major content/service providers are publishing grades they give to ISPs about the quality of delivery of their content. The goal is to inform customers about the “best” ISPs. But this could be an incentive for, or even a pressure on, ISPs to differentiate service and provide a better quality to those big content providers in order to be more attractive. Instead of the traditional vision of ISPs pressing content providers, we face here the opposite situation, still possibly at the expense of small content providers though. We design in [48] a model describing the various actors and their strategies, analyzes it using non-cooperative game theory tools, and quantifies the impact of those advertised grades with respect to the situation where no grade is published. We illustrate that a non-neutral behavior, differentiating traffic, is not leading to a desirable situation.

Sponsored data. With wireless sponsored data, a third party, content or service provider, can pay for some of your data traffic so that it is not counted in your plan's monthly cap. This type of behavior is currently under scrutiny, with telecommunication regulators wondering if it could be applied to prevent competitors from entering the market, and what the impact on all telecommunication actors can be. To answer those questions, we design and analyze in [69] a model where a Content Provider (CP) can choose the proportion of data to sponsor and a level of advertisement to get a return on investment, and several Internet Service Providers (ISPs) in competition. We distinguish three scenarios: no sponsoring, the same sponsoring to all users, and a different sponsoring depending on the ISP you have subscribed to. This last possibility may particularly be considered an infringement of the network neutrality principle. We see that sponsoring can be beneficial to users and ISPs, especially with identical sponsoring. We also discuss the impact of zero-rating where an ISP offers free data to a CP to attract more customers, of and vertical integration where a CP and an ISP are the same company.

Online platforms and search engines. The search neutrality debate is about whether search engines should or should not be allowed to uprank certain results among the organic content matching a query. This debate is related to that of network neutrality, which focuses on whether all bytes being transmitted through the Internet should be treated equally. In a previous paper, we had formulated a model that formalizes this question and characterized an optimal ranking policy for a search engine. The model relies on the trade-off between short-term revenues, captured by the benefits of highly-paying results, and long-term revenues which can increase by providing users with more relevant results to minimize churn. In [21], we apply that model to investigate the relations between search neutrality and innovation. We illustrate through a simple setting and computer simulations that a revenue-maximizing search engine may indeed deter innovation at the content level. Our simple setting obviously simplifies reality, but this has the advantage of providing better insights on how optimization by some actors impacts other actors.

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 impression or per click. But advertisers can be both sensitive to being viewed (brand awareness effect) and being clicked (conversion into sales). In [23], we generalize the auction mechanism by including both pricing components: the 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 extracted from 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.

7.4. Monte Carlo

Participants: Bruno Tuffin, Gerardo Rubino, Pierre L'Ecuyer

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. We have published an introduction to Monte Carlo methods on Insterstices, including animations https://interstices.info/jcms/int_69164/la-simulation-de-monte-carlo.

Rare event simulation. The mean time to failure (MTTF) of a stochastic system is often estimated by simulation. One natural estimator, which we call the direct estimator, simply averages independent and identically distributed copies of simulated times to failure. When the system is regenerative, an alternative approach is based on a ratio representation of the MTTF. The purpose of [42] is to compare the two estimators. We first analyze them in the setting of crude simulation (i.e., no importance sampling), showing that they are actually asymptotically identical in a rare-event context. The two crude estimators are inefficient in different but closely related ways: the direct estimator requires a large computational time because times to failure often include many transitions, whereas the ratio estimator entails estimating a rare-event probability. We then

discuss the two approaches when employing importance sampling; for highly reliable Markovian systems, we show that using a ratio estimator is advised.

Another problem studied in [40] is the estimation of the tail of the distribution of the sum of correlated log-normal random variables. While a number of theoretically efficient estimators have been proposed for this setting, using a few numerical examples we illustrate that these published proposals may not always be useful in practical simulations. As a remedy to this defect, we propose a new estimator and we demonstrate that, not only is our novel estimator theoretically efficient, but, more importantly, its practical performance is significantly better than that of its competitors.

Random variable generation. Random number generators were invented before there were symbols for writing numbers, and long before mechanical and electronic computers. All major civilizations through the ages found the urge to make random selections, for various reasons. Today, random number generators, particularly on computers, are an important (although often hidden) ingredient in human activity. In the invited paper [32], we give a historical account on the design, implementation, and testing of uniform random number generators used for simulation.

We study in [68] the lattice structure of random number generators of the specific MIXMAX family, a class of matrix linear congruential generators that produce a vector of random numbers at each step. These generators were initially proposed and justified as close approximations to certain ergodic dynamical systems having the Kolmogorov K-mixing property, which implies a chaotic (fast-mixing) behavior. But for a K-mixing system, the matrix must have irrational entries, whereas for the MIXMAX it has only integer entries. As a result, the MIXMAX has a lattice structure just like linear congruential and multiple recursive generators. We study this lattice structure for vectors of successive and non-successive output values in various dimensions. We show in particular that for coordinates at specific lags not too far apart, in three dimensions, all the nonzero points lie in only two hyperplanes. This is reminiscent of the behavior of lagged-Fibonacci and AWC/SWB generators. And even if we skip the output coordinates involved in this bad structure, other highly structured projections often remain, depending on the choice of parameters.

Quasi-Monte Carlo (QMC). In [5], which appeared in 2017, we survey basic ideas and results on randomized quasi-Monte Carlo (RQMC) methods, discuss their practical aspects, and give numerical illustrations. RQMC can improve accuracy compared with standard Monte Carlo (MC) when estimating an integral interpreted as a mathematical expectation. RQMC estimators are unbiased and their variance converges at a faster rate (under certain conditions) than MC estimators, as a function of the sample size. Variants of RQMC also work for the simulation of Markov chains, for function approximation and optimization, for solving partial differential equations, etc. In this introductory survey, we look at how RQMC point sets and sequences are constructed, how we measure their uniformity, why they can work for high-dimensional integrals, and how can they work when simulating Markov chains over a large number of steps.

General presentations. Finally, in two general presentations, we described state-of-the-art technologies available to deal with rare events by means of Monte Carlo techniques, including several methods produced inside Dionysos. In the tutorial [33], we gave an overview of the field, with a focus on dependability analysis applications. The keynote [36] described specific procedures taken from our monograph [72], that were adapted to the needs of the micro-simulation community.

7.5. Wireless Networks

Participants: Yue Li, Imad Alawe, Quang Pham, Patrick Maillé, Yassine Hadjadj-Aoul, César Viho, Gerardo Rubino

Mobile wireless networks' improvements. Software Defined Networking (SDN) is one of the key enablers for evolving mobile network architecture towards 5G. SDN involves the separation of control and data plane functions, which leads, in the context of 5G, to consider the separation of the control and data plane functions of the different gateways of the Evolved Packet Core (EPC), namely Serving and Packet data Gateways (S and P-GW). Indeed, the envisioned solutions propose to separate the S/P-GW into two entities: the S/P-GW-C, which integrates the control plane functions and the S/P-GW-U that handles the User Equipment (UE)

data plane traffic. There are two major approaches to create and update user plane forwarding rules for such a partition: (i) considering an SDN controller for the S/P-GW-C (SDNEPC) or (ii) using a direct specific interface to control the S/P-GW-U (enhancedEPC). In [38], we evaluate, using a testbed, those two visions against the classical virtual EPC (vEPC), where all the elements of the EPC are virtualized. Besides evaluating the capacity of the vEPC to manage and scale to UE requests, we compare the performances of the solutions in terms of the time needed to create the user data plane. The obtained results allow drawing several remarks, which may help to dimension the vEPC's components as well as to improve the S/P-GW-U management procedure.

One of the requirements of 5G is to support a massive number of connected devices, considering many use-cases such as IoT and massive Machine Type Communication (MTC). While this represents an interesting opportunity for operators to grow their business, it will need new mechanisms to scale and manage the envisioned high number of devices and their generated traffic. Particularity, the signaling traffic, which will overload the 5G core Network Function (NF) in charge of authentication and mobility, namely Access and Mobility Management Function (AMF). The objective of [37] is to provide an algorithm based on Control Theory allowing: (i) to equilibrate the load on the AMF instances in order to maintain an optimal response time with limited computing latency; (ii) to scale out or in the AMF instance (using NFV techniques) depending on the network load to save energy and avoid wasting resources. Obtained results indicate the superiority of our algorithm in ensuring fair load balancing while scaling dynamically with the traffic load. In [64] we are going further by using new advances on machine learning, and more specifically Recurrent Neural Networks (RNN), to predict accurately the arrival traffic pattern of devices. The main objective of the proposed approach is to early react to congestion by pro-actively scaling the AMF VNF in a way to absorb such congestion while respecting the traffic constraints.

Energy consumption improvements. Recently in cellular networks, the focus has been moved to seeking ways to increase the energy efficiency by better adapting to the existing users behaviors. In [17], we are going a step further in studying a new type of disruptive service by trying to answer the question "What are the potential energy efficiency gains if some of the users are willing to tolerate delays?". We present an analytical model of the energy usage of LTE base stations, which provides lower bounds of the possible energy gains under a decentralized, noncooperative setup. The model is analyzed in six different scenarios (such as micro-macro cell interaction and coverage redundancy) for varying traffic and user-tolerable delays. We show that it is possible to reduce the power consumption by up to 30%.

Computation offloading in mobile network. Mobile edge computing (MEC) emerges as a promising paradigm that extends the cloud computing to the edge of pervasive radio access networks, in near vicinity to mobile users, reducing drastically the latency of end-to-end access to computing resources. Moreover, MEC enables the access to up-to-date information on users' network quality via the radio network information service (RNIS) application programming interface (API), allowing to build novel applications tailored to users' context. In [25] and [49], we present a novel framework for offloading computation tasks, from a user device to a server hosted in the mobile edge (ME) with highest CPU availability. Besides taking advantage of the proximity of the MEC server, the main innovation of the proposed solution is to rely on the RNIS API to drive the user equipment (UE) decision to offload or not computing tasks for a given application. The contributions are twofold. First, we propose the design of an application hosted in the ME, which estimates the current value of the round trip time (RTT) between the UE and the ME, according to radio quality indicators available through RNIS API, and provides it to the UE. Second, we present a novel computation algorithm which, based on the estimated RTT coupled with other parameters (e.g., energy consumption), decide when to offload UE's applications computing tasks to the MEC server. The effectiveness of the proposed framework is demonstrated via testbed experiments featuring a face recognition application.

Services improvement in wireless heterogeneous networks. With the rapid growth of HTTP-based Adaptive Streaming (HAS) multimedia video services on the Internet, improving the Quality of Experience (QoE) of video delivery will be highly requested in wireless heterogeneous networks. Various access technologies such as 3G/LTE and Wi-Fi with overlapping coverage is the main characteristic of network heterogeneity. Since contemporary mobile devices are usually equipped with multiple radio interfaces, mobile users are enabled to

utilize multiple access links simultaneously for additional capacity or reliability. However, network and video quality selection can have notable impact on the QoE of DASH clients facing the video service's requirements, the wireless channel profiles and the costs of the different links. In this context, the emerging Multi-access Edge Computing (MEC) standard gives new opportunities to improve DASH performance, by moving IT and cloud computing capabilities down to the edge of the mobile network. In [45], we propose a MEC-assisted architecture for improving the performance of DASH-based streaming, a standard implementation of a HAS framework in wireless heterogeneous networks. With the proposed algorithm running as a MEC service, the overall QoE and fairness of DASH clients are improved in a real time manner in case of network congestion.

QoE aware routing in wireless networks. This year we continued our research on QoE-based optimization routing for wireless mesh networks. The difficulties of the problem are analyzed and centralized and decentralized algorithms are proposed. The quality of the solution, the computational complexity of the proposed algorithm, and the fairness are our main concerns. Several centralized approximation algorithms have been already proposed in order to address the complexity and the quality of possible solutions. This year, we focused mainly on distributed algorithm to complement of the existing centralized algorithms. We propose decentralized heuristic algorithms based on the well-known Optimized Link-State Routing (OLSR) protocol. Control packets of OLSR are modified so as to be able to convey QoE-related information. The routing algorithm chooses the paths heuristically. After that, we studied message passing algorithms in order to find near optimal routing solutions in cooperative distributed networks. These algorithms have been published in [27], [13].

Sensors networks. In the literature, it is common to consider that sensor nodes in a clustered-based event-driven Wireless Sensor Network (WSN) use a Carrier Sense Multiple Access (CSMA) protocol with a fixed transmission probability to control data transmission. However, due to the highly variable environment in these networks, a fixed transmission probability may lead to a significant amount of extra energy consumption. In view of this, three different transmission probability strategies for event-driven WSNs were studied in [51]: the optimal one, the "fixed" approach and a third "adaptive" method. As expected, the optimum strategy achieves the best results in terms of energy consumption but its implementation in a practical system is not feasible. The commonly used fixed transmission strategy (the probability for any node to attempt transmission is a constant) is the simplest approach but it does not adapt to changes in the system's conditions and achieves the worst performance. In the paper, we find that our proposed adaptive transmission strategy, where that probability is changed depending on specific conditions and in a very precise way, is pretty easy to implement and achieves results very close to the optimal method. The three strategies are analyzed in terms of energy consumption but also regarding the cluster formation latency. In [28], we also investigate cluster head selection schemes. Specifically, we consider two intelligent schemes based on the fuzzy C-means and k-medoids algorithms, and a random selection with no intelligence. We show that the use of intelligent schemes greatly improves the performance of the system, but their use entails higher complexity and some selection delay. The main performance metrics considered in this work are energy consumption, successful transmission probability and cluster formation latency. As an additional feature of this work, we study the effect of errors in the wireless channel and the impact on the performance of the system under the different considered transmission probability schemes.

Transmission delay, throughput and energy are also important criteria to consider in wireless sensor networks (WSNs). The IEEE 802.15.4 standard was conceived with the objective of reducing resource's consumption in both WSNs and Personal Area Networks (WPANs). In such networks, the slotted CSMA/CA still occupies a prominent place as a channel control access mechanism with its inherent simplicity and reduced complexity. In [26], we propose to introduce a network allocation vector (NAV) to reduce energy consumption and collisions in IEEE 802.15.4 networks. A Markov chain-based analytical model of the fragmentation mechanism, in a saturated traffic, is given as well as a model of the energy consumption using the NAV mechanism. The obtained results show that the fragmentation technique improves at the same time the throughput, the access delay and the bandwidth occupation. They also show that using the NAV allows reducing significantly the energy consumption when applying the fragmentation technique in slotted CSMA/CA under saturated traffic conditions.

7.6. Optical Networks

Participants: Nicolás Jara, Gerardo Rubino

The rapid increase in demand for bandwidth in communication networks has caused a growth in the use of technologies based on WDM optical infrastructures. Nevertheless, in this last decade many researchers have recognized a "Capacity Crunch" associated with this technology, a transmission capacity limit on optical fibers, that is close to be reached pretty soon. This situation claims for an evolution on the currently used WDM optical architectures, in order to satisfy this relentless exponential growth in bandwidth demand. Following this trend, research started to examine in some detail specific aspects of the present functioning, and in particular, the way these networks are operated. Currently, optical networks are operated statically, but this is known to be inefficient in the usage of network resources, and with the previously mentioned upcoming risk of capacity collapse, it is of pressing matter to upgrade it. To this purpose, several proposals have been addressed and researched so far. Among these solutions, dynamic optical networks is the one closest to be implemented, but it has not been considered yet since the network cost savings are not enough to convince enterprises. This has been the focus of our research effort in the area.

The design of dynamic optical networks decomposes into different tasks, where the engineers must basically organize the way the main system's resources are used, minimizing the design and operation costs and respecting critical performance constraints. These tasks must guarantee certain level of quality of service (QoS) pre-established in the Service Level Agreement. In order to provide a proper quality of service measurement, we propose a new fast and accurate analytical method to evaluate the blocking probability that is at the heart of the path toward solving all the mentioned design problems. Blocking probability is the main QoS metric considered in the field. This work has been done in [20], where an analytical procedure has been proposed that combines efficiency and accuracy.

Next, the different tasks that must be addressed to find a good global design have been addressed in [19]. These are: which wavelength is going to be used by each user (the Wavelength Assignment Problem), how many wavelengths will be needed on each network link (the Wavelength Dimensioning Problem), and which set of paths enabling each network user to transmit (known as the Routing Problem) are to be established in order to minimize costs and to deal with link failures when the network is operating (this is the Fault Tolerance Problem). Two types of innovations and presented in this last paper. First, each of the problems receives a solution shown to be highly efficient. Second, and this is also new, we solve all the design problems simultaneously, using a single global algorithm (the usual way is to isolate them and to solve them one at a time, in a specific order). This work may provide a strategy to finally achieve sufficient cost savings, and thus, to contribute to make the decision to migrate from static to dynamic resource allocation easier. A preliminary version of a part of these results was presented previously in [44].

7.7. Future networks and architectures

Participants: Jean-Michel Sanner, Hamza Ben Ammar, Louiza Yala, Yassine Hadjadj-Aoul, Gerardo Rubino

SDN and NFV placement. Mastering the increasing complexity of current and future networks, while reducing the operational and investments costs, is one of the major challenges faced by network operators (NOs). This explains in large part the recent enthusiasm of NOs towards Software Defined Networking (SDN) and Network Function Virtualization (NFV). Indeed, on the one hand, SDN makes it possible to get rid of the control plane distribution complexity, by centralizing it logically, while allowing its programmability. On the other hand, the NFV allows virtualizing the network functions, which considerably facilitates the deployment and the orchestration of the network resources. Providing a carrier grade network involves, however, several requirements such as providing a robust network meeting the constraints of the supported services. In order to achieve this objective, it is clearly necessary to scale network functions while placing them strategically in a way to guarantee the system's responsiveness.

The placement in TelCo networks are generally multi-objective and multi-constrained problems. The solutions proposed in the literature usually model the placement problem by providing a mixed integer linear program (MILP). Their performances are, however, quickly limited for large sized networks, due to the significant increase in the computational delays. In order to avoid the inherent complexity of optimal approaches and the

lack of flexibility of heuristics, we propose in [54] a genetic algorithm designed from the NSGA II framework that aims to deal with the controller placement problem. Genetic algorithms can be both multi-objective, multi-constraints and can be designed to be implemented in parallel. They constitute a real opportunity to find good solutions to this category of problems. Furthermore, the proposed algorithm can be easily adapted to manage dynamic placements scenarios. In [55], our main focus were devoted to maximize the clusters average connectivity and to balance the control's load between clusters, in a way to improve the networks' reliability.

We focus, in [60], on the problem of optimal computing resource allocation and placement for the provision of a virtualized Content Delivery Network (CDN) service over a telecom operator's Network Functions Virtualization (NFV) infrastructure. Starting from a Quality of Experience (QoE)-driven decision on the necessary amount of CPU resources to allocate in order to satisfy a virtual CDN deployment request with QoE guarantees, we address the problem of distributing these resources to virtual machines and placing the latter to physical hosts, optimizing for the conflicting objectives of management cost and service availability, while respecting physical capacity, availability and cost constraints. We present a multi-objective optimization problem formulation, and provide efficient algorithms to solve it by relaxing some of the original problem's assumptions. Numerical results demonstrate how our solutions address the trade-off between service availability and cost, and show the benefits of our approach compared with resource placement algorithms which do not take this trade-off into account.

Real-time NFV placement in edge cloud. Sometimes, the placement of NFV can not be planned in advance and therefore requires real-time placement as requests arrive. The placement is particularly challenging with the recent development of geographically distributed mini data centers, also referred to as cloudlets, at the edge of the network (i.e., typically at Points of Presence (PoPs) level). These edge data centers have rather small capacities in terms of storage, computing and networking resources, when compared with the huge centralized data centers deployed today.

All these radical changes in NOs' infrastructures raise many new issues (especially in terms of resource allocation), which so far have not been considered in the cloud literature. Traditionally, resources in cloud platforms are considered as to be infinite and request blocking is most of the time ignored when evaluating resources' allocation algorithms, precisely because of this infinite capacity assumption. However, if we assume that the NO's infrastructure will very likely be composed of small data centers with limited capacities, and deployed at the edge of network, the congestion of such a system may occur, notably if the demand is sufficiently high and exceeds what the infrastructure can handle at a given time.

We proposed in [57] an analytical model for the blocking analysis in a multidimensional cloud system, which was validated using discrete events' simulations. Besides, we conducted a comparative analysis of the most popular placement's strategies. The proposed model, as well as the comparative study, reveal practical insights into the performance evaluation of resource allocation and capacity planning for distributed edge cloud with limited capacities.

In [58] we set design principles of future distributed edge clouds in order to meet application requirements. We precisely introduce a costless distributed resource allocation algorithm, named *CLOSE*, which considers local information only. We compare via simulations the performance of *CLOSE* against those obtained by using mechanisms proposed in the literature, notably the Tricircle project within OpenStack. It turns out that the proposed distributed algorithm yields better performance while requiring less overhead.

In the context of the Open Network Automation Platform (ONAP), we develop in [56] a resource allocation strategy for deploying Virtualized Network Functions (VNFs) on distributed data centers. For this purpose, we rely on a three-level data center hierarchy exploiting co-location facilities available within Main and Core Central Offices. We precisely propose an active VNFs' placement strategy, which dynamically offloads requests on the basis of the load observed within a data center. We compare via simulations the performance of the proposed solution against mechanisms so far proposed in the literature, notably the centralized approach of the multi-site project within OpenStack, currently adopted by ONAP. Our algorithm yields better performance in terms of both data center occupancy and overhead. Furthermore, it allows extending the applicability of ONAP in the context of distributed cloud, without requiring any modification.

Content Centric Networking. Content-Centric Networking (CCN) has been proposed to address the challenges raised by the Internet usage evolution over the last years. One key feature provided by CCN to improve the efficiency of content delivery is the in-network caching, which has major impact on the system performance. In order to improve caching effectiveness in such systems, the study of the functioning of CCN in-network storage must go deeper. In [39], we propose MACS, a Markov chain-based Approximation of CCN caching Systems. We start initially by modeling a single cache node. Then, we extend our model to the case of multiple nodes. A closed-form expression is then derived to define the cache hit probability of each content in the caching system. We compare the results of MACS to those obtained with simulations. The conducted experiments show clearly the accuracy of our model in estimating the cache hit performance of the system.

In [16], we present the design and implementation of a Content-Delivery-Network-as-a-Service (CDNaaS) architecture, which allows a telecom operator to open up its cloud infrastructure for content providers to deploy virtual CDN instances on demand, at regions where the operator has presence. Using northbound REST APIs, content providers can express performance requirements and demand specifications, which are translated into an appropriate service placement on the underlying cloud substrate. Our architecture is extensible, supporting various different CDN flavors, and, in turn, different schemes for cloud resource allocation and management. In order to decide on the latter in an optimal manner from an infrastructure cost and a service quality perspective, knowledge of the performance capabilities of the underlying technologies and computing resources is critical. Therefore, to gain insight which can be applied to the design of such mechanisms, but also with further implications on service pricing and SLA design, we carry out a measurement campaign to evaluate the capabilities of key enabling technologies for CDNaaS provision. In particular, we focus on virtualization and containerization technologies for implementing virtual CDN functions to deliver a generic HTTP service, as well as an HTTP video streaming one, empirically capturing the relationship between performance and service workload, both from a system operator and a user-centric viewpoints.

New tools for network design. In the efforts for designing future networks' topologies, the inclusion of dependability aspects has been recently enriched with finer criteria, and one relatively new family of metrics consider diameter-constrained parameters that capture more accurately reliability aspects of communication infrastructures. This is done by taking into account not only connectivity properties but also delays when nodes are connected. Paper [15] deals with factorization theory in diameter-constrained reliability, when terminal nodes are further required to be connected by d hops or fewer (d is a given strictly positive parameter of the metric, called its diameter). This metric was defined in 2001, inspired by delay-sensitive applications in telecommunications. Factorization theory is fundamental for classical network reliability evaluation, and today it is a mature area. However, its extension to the diameter-constrained context requires at least the recognition of irrelevant links, which is an open problem. In this paper, irrelevant links are efficiently determined in the most used case, where we consider the communication between a given pair of nodes in the network. The article also proposes a Factoring algorithm that includes the way series-parallels substructures can be handled.

Quality of Experience activities. We continue to develop tools for Quality of Experience assessment, and applications of this quantitative evaluation.

Predicting time series. For the future of the PSQA project, we intend to integrate the capability of *predicting* the Perceptual Quality and not only evaluating its current value. With this goal in mind, we explored this year the idea of combining a Reservoir Computing architecture (whose good performances have been reported many times, when used to predict sequences of numbers or of vectors) with Recurrent Random Neural Networks, that belong to a class of Neural Networks that have some nice properties. Both have been very successful in many applications. In [29] we propose a new model belonging to the first class, taking the structure of the second for its dynamics. The new model is called Echo State Queuing Network. The paper positions the model in the global Machine Learning area, and provides examples of its use and performances. We show on largely used benchmarks that it is a very accurate tool, and we illustrate how it compares with standard Reservoir Computing models. In [31] we presented some preliminary results to the Random Neural Network community.

QoE and P2P design. In [30] we describe a Peer-to-Peer (P2P) network that was designed to support Video on Demand (VoD) services. The network is based on a video-file sharing mechanism that classifies peers

according to the window (segment of the file) that they are downloading. This classification easily allows identifying peers that are able to share windows among them, so one of our major contributions is the definition of a mechanism that could be implemented to efficiently distribute video content in future 5G networks. Considering that cooperation among peers can be insufficient to guarantee an appropriate system performance, we also propose that this network must be assisted by upload bandwidth coming from servers; since these resources represent an extra cost to the service provider, especially in mobile networks, we complement our work by defining a scheme that efficiently allocates them only to those peers that are in windows with resources scarcity (we called it *prioritized windows distribution scheme*). On the basis of a fluid model and a Markov chain, we also develop a methodology that allows us to select the system parameters values (e.g., windows sizes or minimum servers upload bandwidth) that satisfy a set of Quality of Experience (QoE) parameters.

DIVERSE Project-Team

7. New Results

7.1. Results on Variability modeling and management

7.1.1. Variability and testing.

Many approaches for testing configurable software systems start from the same assumption: it is impossible to test all configurations. This motivated the definition of variability-aware abstractions and sampling techniques to cope with large configuration spaces. Yet, there is no theoretical barrier that prevents the exhaustive testing of all configurations by simply enumerating them, if the effort required to do so remains acceptable. Not only this: we believe there is lots to be learned by systematically and exhaustively testing a configurable system. We report on the first ever endeavor to test all possible configurations of an industry-strength, open source configurable software system, JHipster, a popular code generator for web applications. We built a testing scaffold for the 26,000+ configurations of JHipster using a cluster of 80 machines during 4 nights for a total of 4,376 hours (182 days) CPU time. We find that 35.70% configurations fail and we identify the feature interactions that cause the errors. We show that sampling testing strategies (like dissimilarity and 2-wise) (1) are more effective to find faults than the 12 default configurations used in the JHipster continuous integration; (2) can be too costly and exceed the available testing budget. We cross this quantitative analysis with the qualitative assessment of JHipster's lead developers. Additional resources: preliminary effort on JHipster [32], https://arxiv.org/abs/1710.07980https://github.com/axel-halin/Thesis-JHipster/

7.1.2. Variability and teaching.

Software Product Line (SPL) engineering has emerged to provide the means to efficiently model, produce, and maintain multiple similar software variants, exploiting their common properties, and managing their variabilities (differences). With over two decades of existence, the community of SPL researchers and practitioners is thriving as can be attested by the extensive research output and the numerous successful industrial projects. Education has a key role to support the next generation of practitioners to build highly complex, variability-intensive systems. Yet, it is unclear how the concepts of variability and SPLs are taught, what are the possible missing gaps and difficulties faced, what are the benefits, or what is the material available. Also, it remains unclear whether scholars teach what is actually needed by industry. We report on three initiatives we have conducted with scholars, educators, industry practitioners, and students to further understand the connection between SPLs and education, i.e., an online survey on teaching SPLs we performed with 35 scholars, another survey on learning SPLs we conducted with 25 students, as well as two workshops held at the International Software Product Line Conference in 2014 and 2015 with both researchers and industry practitioners participating. We build upon the two surveys and the workshops to derive recommendations for educators to continue improving the state of practice of teaching SPLs, aimed at both individual educators as well as the wider community. Finally, we are developing and maintaining a repository for teaching SPLs and variability. Additional resources: https://teaching.variability.io

7.1.3. Variability and constraint solving.

Array constraints are essential for handling data structures in automated reasoning and software verification. Unfortunately, the use of a typical finite domain (FD) solver based on local consistency-based filtering has strong limitations when constraints on indexes are combined with constraints on array elements and size. This work proposes an efficient and complete FD-solving technique for extended constraints over (possibly unbounded) arrays. We describe a simple but particularly powerful transformation for building an equisatisfiable formula that can be efficiently solved using standard FD reasoning over arrays, even in the unbounded case. Experiments show that the proposed solver significantly outperforms FD solvers, and successfully competes with the best SMT-solvers [38]. This work is not directly related to variability and SPL. But it contributes to DiverSE's attempts to connect artificial intelligence techniques to software variability engineering, in which constraint solving or machine learning are typically applied.

7.1.4. Variability and machine learning (performance specialization of variability-intensive systems).

We propose the use of a machine learning approach to infer variability constraints from an oracle that is able to assess whether a given configuration is correct. We propose an automated procedure to randomly generate configurations, classify them according to the oracle, and synthesize cross-tree constraints. Specifically, based on an oracle (e.g. a runtime test) that tells us whether a given configuration meets the requirements (e.g. speed or memory footprint), we leverage machine learning to retrofit the acquired knowledge into a variability model of the system that can be used to automatically specialize the configurable system. We validate our approach on a set of well-known configurable software systems (Apache server, x264, etc.) Our results show that, for many different kinds of objectives and performance qualities, the approach has interesting accuracy, precision and recall after a learning stage based on a relative small number of random samples [43]. Additional resources: https://learningconstraints.github.io and VaryVary ANR project

7.1.5. Variability and machine learning (learning contextual variability models).

Modeling how contextual factors relate to a software system's configuration space is usually a manual, error-prone task that depends highly on expert knowledge. Machine-learning techniques can automatically predict the acceptable software configurations for a given context. Such an approach executes and observes a sample of software configurations within a sample of contexts. It then learns what factors of each context will likely discard or activate some of the software's features. This lets developers and product managers automatically extract the rules that specialize highly configurable systems for specific contexts [27] Additional resources: https://learningconstraints.github.io and VaryVary ANR project

We are currently exploring the use of machine learning for variability-intensive systems in the context of VaryVary ANR project (see also VaryLaTeX [28]).

7.2. Results on Software Language Engineering

7.2.1. On Language Interfaces

Complex systems are developed by teams of experts from multiple domains , who can be liberated from becoming programming experts through domain-specific languages (DSLs). The implementation of the different concerns of DSLs (including syntaxes and semantics) is now well-established and supported by various languages workbenches. However, the various services associated to a DSL (e.g., editors, model checker, debugger or composition operators) are still directly based on its implementation. Moreover, while most of the services crosscut the different DSL concerns, they only require specific information on each. Consequently, this prevents the reuse of services among related DSLs, and increases the complexity of service implementation. Leveraging the time-honored concept of interface in software engineering, we discuss in [40] the benefits of language interfaces in the context of software language engineering. In particular, we elaborate on particular usages that address current challenges in language development.

7.2.2. Revisiting Visitors for Modular Extension of Executable DSMLs

Executable Domain-Specific Modeling Languages (xDSMLs) are typically defined by metamodels that specify their abstract syntax, and model interpreters or compilers that define their execution semantics. To face the proliferation of xDSMLs in many domains, it is important to provide language engineering facilities for opportunistic reuse, extension, and customization of existing xDSMLs to ease the definition of new ones. Current approaches to language reuse either require to anticipate reuse, make use of advanced features that are not widely available in programming languages, or are not directly applicable to metamodel-based xDSMLs. In [35], we propose a new language implementation pattern, named REVISITOR, that enables independent extensibility of the syntax and semantics of metamodel-based xDSMLs with incremental compilation and without anticipation. We seamlessly implement our approach alongside the compilation chain of the Eclipse Modeling Framework, thereby demonstrating that it is directly and broadly applicable in various modeling environments. We show how it can be employed to incrementally extend both the syntax and semantics of the fUML language without requiring anticipation or re-compilation of existing code, and with acceptable performance penalty compared to classical handmade visitors.

7.2.3. Advanced and efficient execution trace management for executable domain-specific modeling languages

Executable Domain-Specific Modeling Languages (xDSMLs) enable the application of early dynamic verification and validation (V&V) techniques for behavioral models. At the core of such techniques, execution traces are used to represent the evolution of models during their execution. In order to construct execution traces for any xDSML, generic trace metamodels can be used. Yet, regarding trace manipulations, generic trace metamodels lack efficiency in time because of their sequential structure, efficiency in memory because they capture superfluous data, and usability because of their conceptual gap with the considered xDSML. We contributed in [22] a novel generative approach that defines a multidimensional and domain-specific trace metamodel enabling the construction and manipulation of execution traces for models conforming to a given xDSML. Efficiency in time is improved by providing a variety of navigation paths within traces, while usability and memory are improved by narrowing the scope of trace metamodels to fit the considered xDSML. We evaluated our approach by generating a trace metamodel for fUML and using it for semantic differencing, which is an important V&V technique in the realm of model evolution. Results show a significant performance improvement and simplification of the semantic differencing rules as compared to the usage of a generic trace metamodel.

7.2.4. Omniscient Debugging for Executable DSLs

Omniscient debugging is a promising technique that relies on execution traces to enable free traversal of the states reached by a model (or program) during an execution. While a few General-Purpose Languages (GPLs) already have support for omniscient debugging, developing such a complex tool for any executable Domain Specific Language (DSL) remains a challenging and error prone task. A generic solution must: support a wide range of executable DSLs independently of the metaprogramming approaches used for implementing their semantics; be efficient for good responsiveness. Our contribution in [21] relies on a generic omniscient debugger supported by efficient generic trace management facilities. To support a wide range of executable DSLs, the debugger provides a common set of debugging facilities, and is based on a pattern to define runtime services independently of metaprogramming approaches. Results show that our debugger can be used with various executable DSLs implemented with different metaprogramming approaches. As compared to a solution that copies the model at each step, it is on average six times more efficient in memory, and at least 2.2 faster when exploring past execution states, while only slowing down the execution 1.6 times on average.

7.2.5. Reverse Engineering Language Product Lines from Existing DSL Variants

The use of domain-specific languages (DSLs) has become a successful technique in the development of complex systems. Nevertheless, the construction of this type of languages is time-consuming and requires highly-specialized knowledge and skills. An emerging practice to facilitate this task is to enable reuse through the definition of language modules which can be later put together to build up new DSLs. In [26], we propose a reverse-engineering technique to ease-off such a development scenario. Our approach receives a set of DSL variants which are used to automatically recover a language modular design and to synthesize the corresponding variability models. The validation is performed in a project involving industrial partners that required three different variants of a DSL for finite state machines. This validation shows that our approach is able to correctly identify commonalities and variability.

7.2.6. Software Language Engineering for Virtual Reality Software Development

Due to the nature of Virtual Reality (VR) research, conducting experiments in order to validate the researcher's hypotheses is a must. However, the development of such experiments is a tedious and time-consuming task. In [48], we propose to make this task easier, more intuitive and faster with a method able to describe and generate the most tedious components of VR experiments. The main objective is to let experiment designers focus on their core tasks: designing , conducting, and reporting experiments. To that end, we applied well-established SLE concepts promoted in DIVERSE to the VR domain to ease the development of VR experiments. More precisely, we propose the use of DSLs to ease the description and generation of VR experiments. An analysis of published VR experiments is used to identify the main properties that characterize VR experiments.

This allowed us to design AGENT (Automatic Generation of ExperimeNtal proTocol runtime), a DSL for specifying and generating experimental protocol runtimes. We demonstrated the feasibility of our approach by using AGENT on two experiments published in the VRST'16 proceedings.

7.2.7. Create and Play your Pac-Man Game with the GEMOC Studio

Executable Domain-Specific Languages (DSLs) are used for defining the behaviors of systems. In particular, the operational semantics of such DSLs may define how conforming models react to stimuli from their environment. This commonly requires adapting the semantics to define both the possible domain-level stimuli, and their handling during the execution. However, manually adapting the semantics for such cross-cutting concern is a complex and error-prone task. In , we demonstrate a tool addressing this problem by allowing the augmentation of operational semantics for handling stimuli, and by automatically generating a complete behavioral language interface from this augmentation. At runtime, this interface can receive stimuli sent to models, and can safely handle them by automatically interrupting the execution flow. This tool is an extension to the GEMOC Studio, a language and modeling workbench for executable DSLs We demonstrate how it can be used to implement a Pac-Man DSL enabling to create and play Pac-Man games.

7.3. Results on Heterogeneous and dynamic software architectures

We have selected three main contributions for DIVERSE's research axis #4: one is in the field of runtime management, while the two others one are in the field of Privacy and Security.

7.3.1. Verifying the configuration of Virtualized Network Functions in Software Defined Networks

In Kevoree, one of the goal is to work on the shipping pases in which we aim at making deployment, and the reconfiguration simple and accessible to the whole team. This year we work to include the capacity to manage network configuration when reconfiguring application stack. In this context, the deployment of modular virtual network functions (VNFs) in software defined infrastructures (SDI) enables cloud and network providers to deploy integrated network services across different resource domains. It leads to a large interleaving between network configuration through software defined network controllers and VNF deployment within this network. Most of the configuration management tools and network orchestrator used to deploy VNF lack of an abstraction to express Assume-Guarantee contracts between the VNF and the SDN configuration. Consequently, VNF deployment can be inconsistent with network configurations.

Contribution. To tackle this challenge, in this work [41], we develop an approach to check the consistency between the VNF description described from a set of structural models and flow-chart models and a proposed deployment on a real SDN infrastructure with its own configuration manager. We illustrate our approach on virtualized Evolved Packet Core function.

Originality. The originality of this work is to propose a model to capture VNF.

Impact. Beyond the scientific originality of this work, the main impacts of this novel approach to check SDN configuration has been to (i) reinforce DIVERSE's visibility in the academic and industrial communities on software components and (ii) to create several research tracks that are currently explored in different projects of the team (B-com PhD thesis and Nokia common labs). This work is being integrated within the Kevoree platform.

7.3.2. Identity Negotiation at Runtime

Authentication delegation is a major function of the modern web. Identity Providers (IdP) acquired a central role by providing this function to other web services. By knowing which web services or web applications access its service, an IdP can violate the end-user privacy by discovering information that the user did not want to share with its IdP. For instance, WebRTC introduces a new field of usage as authentication delegation happens during the call session establishment, between two users. As a result, an IdP can easily discover that Bob has a meeting with Alice. A second issue that increases the privacy violation is the lack of choice for the end-user to select its own IdP. Indeed, on many web-applications, the end-user can only select between a subset of IdPs, in most cases Facebook or Google.

Contribution. This year, we analyze this phenomena [23], in particular why the end-user cannot easily select its preferred IdP, though there exists standards in this field such as OpenID Connect and OAuth 2? To lead this analysis, we conduct three investigations. The first one is a field survey on OAuth 2 and OpenID Connect scope usage by web sites to understand if scopes requested by web-sites could allow for user defined IdPs. The second one tries to understand whether the problem comes from the OAuth 2 protocol or its implementations by IdP. The last one tries to understand if trust relations between websites and IdP could prevent the end user to select its own IdP. Finally, we sketch possible architecture for web browser based identity management, and report on the implementation of a prototype. We also describe our implementation of the WebRTC identity architecture [24]. We adapt OpenID Connect servers to support WebRTC peer to peer authentication and detail the issues and solutions found in the process.

Originality. We observe that although WebRTC allows for the exchange of identity assertion between peers, users lack feedback and control over the other party authentication. To allow identity negotiation during a WebRTC communication setup, we propose an extension to the Session Description Protocol. Our implementation demonstrates current limitations with respect to the current WebRTC specification.

Impact. This work is done with Orange.

7.3.3. Raising Time Awareness in Model-Driven Engineering

The conviction that big data analytics is a key for the success of modern businesses is growing deeper, and the mobilisation of companies into adopting it becomes increasingly important. Big data integration projects enable companies to capture their relevant data, to efficiently store it, turn it into domain knowledge, and finally monetize it. In this context, historical data, also called temporal data, is becoming increasingly available and delivers means to analyse the history of applications, discover temporal patterns, and predict future trends. Despite the fact that most data that today's applications are dealing with is inherently temporal current approaches, methodologies, and environments for developing these applications don't provide sufficient support for handling time. We envision that Model-Driven Engineering (MDE) would be an appropriate ecosystem for a seamless and orthogonal integration of time into domain modeling and processing.

Contribution. This year, we investigate the state-of-the-art in MDE techniques and tools in order to identify the missing bricks for raising time-awareness in MDE and outline research directions in this emerging domain [30].

Originality. We propose an extended context representation for self-adaptive software that integrates the history of planned actions as well as their expected effects over time into the context representations. We demonstrate on a cloud elasticity manager case study that such *temporal action-aware context* leads to improved reasoners while still be highly scalable. This work is original with respect to the state of the art since it provides a way to represent and take into account the impact of reconfiguration actions on a system.

Impact. This work is done through a collaboration with the SnT in Luxembourg and a startup called DataThings, working on domain model representation for various industrial domains.

7.3.4. Collaborations

This year, we had a close and fruitful collaboration with the industrial partners that are involved in the HEADS and Occiware projects, in particular an active interaction with the Tellu company in Norway in the Heads context. Tellu relies on Kevoree and KevoreeJS to build their health management systems. They will be also an active member the new Stamp project led by DIVERSE. We can cite also an active collaboration with Orange Labs through Kevin Corre's joint PhD thesis. Another joint industrial (CIFRE) PhD started in September 2016, and we are also partner in a new starting FUI project. Finally, DIVERSE collaborates with the B-COM IRT (https://b-com.com/en), as one permanent member has a researcher position of one day per week at B-COM and a new joint PhD started in September.

At the academic level we collaborate actively with the Spiral team at Inria Lille (several joint projects), the Tacoma team (with two co-advised PhD students), the Myriad team (1 co-advised PhD student) and we have started two collaborations with the ASAP team.

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, 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

Our work on software diversification explores various ways of adding randomness in program executions: state perturbations that preserve functional correctness [25]; randomizing of web APIs to mitigate browser fingeprinting [33].

Can the execution of software be perturbed without breaking the correctness of the output? In this work [25], we devise a protocol to answer this question from a novel perspective. In an experimental study, we observe that many perturbations do not break the correctness in ten subject programs. We call this phenomenon "correctness attraction". The uniqueness of this protocol is that it considers a systematic exploration of the perturbation space as well as perfect oracles to determine the correctness of the output. To this extent, our findings on the stability of software under execution perturbations have a level of validity that has never been reported before in the scarce related work. A qualitative manual analysis enables us to set up the first taxonomy ever of the reasons behind correctness attraction.

The rich programming interfaces (APIs) provided by web browsers can be diverted to collect a browser fingerprint. A small number of queries on these interfaces are sufficient to build a fingerprint that is statistically unique and very stable over time. Consequently, the fingerprint can be used to track users. Our work [33] aims at mitigating the risk of browser fingerprinting for users privacy by 'breaking' the stability of a fingerprint over time. We add randomness in the computation of selected browser functions, in order to have them deliver slightly different answers for each browsing session. Randomization is possible thanks to the following properties of browsers implementations: (i) some functions have a nondeterministic specification, but a deterministic implementation ; (ii) multimedia functions can be slightly altered without deteriorating user's perception. We present FPRandom, a modified version of Firefox that adds randomness to mitigate the most recent fingerprinting algorithms, namely canvas fingerprinting, AudioContext fingerprinting and the unmasking of browsers through the order of JavaScript properties. We evaluate the effectiveness of FPRandom by testing it against known fingerprinting tests. We also conduct a user study and evaluate the performance overhead of randomization to determine the impact on the user experience.

The other aspect in the area of software diversity is about the statistical analysis of browser fingerprinting on a large industrial dataset [17], [31].

7.4.2. Software testing

Generative software development has paved the way for the creation of multiple code generators and compilers that serve as a basis for automatically generating code to a broad range of software and hardware platforms. With full automatic code generation, the user is able to easily and rapidly synthetize software artifacts for various software platforms. In addition, modern generators (i.e., C compilers) become highly configurable, offering numerous configuration options that the user can use to easily customize the generated code for the target hardware platform. In this context, it is crucial to verify the correct behaviour of code generators. Numerous approaches have been proposed to verify the functional outcome of generated code but few of them evaluate the non-functional properties of automatically generated code, namely the performance and resource usage properties. The thesis of Mohamed Boussaa [16] has addressed this limitation.

KERDATA Project-Team

6. New Results

6.1. Convergence of HPC and Big Data

6.1.1. Týr: Blob-based storage convergence of HPC and Big Data

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

The increasingly growing data sets processed on HPC platforms raise major challenges for the underlying storage layer. A promising alternative to POSIX-I/O-compliant file systems are simpler blobs (binary large objects), or object storage systems. They offer lower overhead, better performance and horizontal scalability at the cost of largely unused features such as file hierarchies or permissions. Similarly, blobs are increasingly considered for replacing distributed file systems for big data analytics or as a base for storage abstractions like key-value stores or time-series databases.

This growing interest from both HPC and Big Data communities towards blob storage naturally fits with the current trend towards HPC and Big Data convergence. In this context, we seek to demonstrate that blob storage indeed constitutes a strong alternative to current storage infrastructures. Additionally, the data model of blob storage is close enough to that of distributed file systems so that this change is largely transparent for the applications running atop them.

In [22] we provide a preliminary evaluation of blob storage in HPC and Big Data contexts. We leverage a series of real-world HPC applications as well as an industry-standard HPC benchmark. We analyze for each of these applications the storage requests sent to the underlying storage system. We discover that over 98% of these storage calls can be directly mapped to the data model offered by blobs. Interestingly, we also note that the remaining calls are using file systems features for convinience rather than by necessity. These calls may consequently be performed as offline pre- or post-processing, or avoided altogether without altering the application.

6.1.2. Modeling elastic storage

Participants: Nathanaël Cheriere, Gabriel Antoniu.

For efficient Big Data processing, efficient resource utilization becomes a major concern as large-scale computing infrastructures such as supercomputers or clouds keep growing in size. Naturally, energy and cost savings can be obtained by reducing idle resources. Malleability, which is the possibility for resource managers to *dynamically* increase or reduce the resources of jobs, appears as a promising means to progress towards this goal.

However, state-of-the-art parallel and distributed file systems have not been designed with malleability in mind. This is mainly due to the supposedly high cost of storage decommission, which is considered to involve expensive data transfers. Nevertheless, as network and storage technologies evolve, old assumptions on potential bottlenecks can be revisited.

In [18], we evaluate the viability of malleability as a design principle for a distributed file system. We specifically model the duration of the decommission operation, for which we obtain a theoretical lower bound. Then we consider HDFS as a use case and we show that our model can explain the measured decommission times.

The existing decommission mechanism of HDFS is good when the network is the bottleneck, but could be accelerated by up to a factor 3 when the storage is the limiting factor. With the highlights provided by our model, we suggest improvements to speed up decommission in HDFS and we discuss open perspectives for the design of efficient malleable distributed file systems.

6.1.3. Eley: Leveraging burst-buffers for efficient Big Data processing on HPC systems

Participants: Orçun Yildiz, Chi Zhou, Shadi Ibrahim.

Burst Buffer is an effective solution for reducing the data transfer time and the I/O interference in HPC systems. Extending Burst Buffers (BBs) to handle Big Data applications is challenging because BBs must account for the large data inputs of Big Data applications and the performance guarantees of HPC applications – which are considered as first-class citizens in HPC systems. Existing BBs focus on only intermediate data of Big Data applications and incur a high performance degradation of both Big Data and HPC applications.In [26], we present *Eley*, a burst buffer solution that helps to accelerate the performance of Big Data applications while guaranteeing the performance of HPC applications. In order to improve the performance of Big Data applications, *Eley* employs a prefetching technique that fetches the input data of these applications to be stored close to computing nodes thus reducing the latency of reading data inputs. Moreover, Eley is equipped with a full delay operator to guarantee the performance of HPC applications – as they are running independently on a HPC system. The experimental results show the effectiveness of *Eley* in obtaining shorter execution time of Big Data applications (shorter map phase) while guaranteeing the performance of HPC applications.

6.2. Scalable data processing on clouds

6.2.1. Low-latency storage for stream processing

Participants: Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu, María Pérez, Radu Tudoran, Stefano Bortoli, Bogdan Nicolae.

We are now witnessing an unprecedented growth of data that needs to be processed at always increasing rates in order to extract valuable insights. Big Data applications are rapidly moving from a batch-oriented execution model to a streaming execution model in order to extract value from the data in real-time. Big Data streaming analytics tools have been developed to cope with the online dimension of data processing: they enable real-time handling of live data sources by means of stateful aggregations (window-based operators). In [21] we design a deduplication method specifically for window-based operators that rely on key-value stores to hold a shared state. Our key finding is that more fine-grained interactions between streaming engines and (key-value) stores (i.e., the data ingest, store, and process interfaces) need to be designed in order to better respond to scenarios that have to overcome memory scarcity.

Moreover, processing live data alone is often not enough: in many cases, such applications need to combine the live data with previously archived data to increase the quality of the extracted insights. Current streaming-oriented runtimes and middlewares are not flexible enough to deal with this trend, as they address ingestion (collection and pre-processing of data streams) and persistent storage (archival of intermediate results) using separate services. This separation often leads to I/O redundancy (e.g., write data twice to disk or transfer data twice over the network) and interference (e.g., I/O bottlenecks when collecting data streams and writing archival data simultaneously). In [20] and [27] we argue for a unified ingestion and storage architecture for streaming data that addresses the aforementioned challenge and we identify a set of constraints and benefits for such a unified model, while highlighting the important architectural aspects required to implement it in real life.

Based on these findings, we are currently developing a low-latency stream storage framework that addresses such critical real-time needs for efficient stream processing, exposing high-performance interfaces for stream ingestion, storage, and processing.

6.2.2. A Performance Evaluation of Apache Kafka in Support of Big Data Streaming Applications

Participants: Paul Le Noac'h, Alexandru Costan.

Stream computing is becoming a more and more popular paradigm as it enables the real-time promise of data analytics. Apache Kafka is currently the most popular framework used to ingest the data streams into the processing platforms. However, how to tune Kafka and how much resources to allocate for it remains a challenge for most users, who now rely mainly on empirical approaches to determine the best parameter settings for their deployments. Our goal in [28] is to make a through evaluation of several configurations and performance metrics of Kafka in order to allow users avoid bottlenecks, reach its full potential and avoid bottlenecks and eventually leverage some good practice for efficient stream processing.

6.2.3. Hot metadata management for geographically distributed workflows

Participants: Luis Eduardo Pineda Morales, Alexandru Costan, Gabriel Antoniu, Ji Liu, Esther Pacitti, Patrick Valduriez, Marta Mattoso.

Large-scale scientific applications are often expressed as scientific workflows (SWfs) that help defining data processing jobs and dependencies between jobs' activities. Several SWfs have huge storage and computation requirements, and so they need to be processed in multiple (cloud-federated) datacenters. It has been shown that efficient metadata handling plays a key role in the performance of computing systems. However, most of this evidence concern only single-site, HPC systems to date. In addition, the efficient scheduling of tasks among different datacenters is critical to the SWf execution. In [19], we present a hybrid distributed model and architecture, using hot metadata (frequently accessed metadata) for efficient SWf scheduling in a multisite cloud. We couple our model with a scientific workflow management system (SWfMS) to validate its applicability to real-life scientific workflows with different scheduling algorithms. We show that the combination of efficient management of hot metadata and scheduling algorithms improves the performance of SWfMS, reducing the execution time of highly parallel jobs up to 64.1 % and that of the whole scientific workflows up to 37.5 %, by avoiding unnecessary cold metadata operations. We also further discuss how to dynamically handle such hot metadata.

6.3. Scalable I/O, storage and in-situ processing in Exascale environments

6.3.1. Extreme-scale logging through application-defined storage

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Applications generating data as logs and seeking to store it as such face hard challenges on HPC platforms. In distributed systems this storage model is key to ensuring fault-tolerance, developing transactional systems or publish-subscribe models. In scientific applications, distributed logs can play many roles such as in-situ visualization of large data streams, centralized collection of telemetry or monitoring events computational steering, data aggregation from array of physical sensors or live data indexing. Distributed shared logs are very difficult to implement on common HPC platforms due to the lack of efficient append operation in the current file-based storage infrastructures. While part of the POSIX standard, this operation has not been the main focus during the development of parallel file systems. While application-specific, custom-built solutions are possible, they require a significant development effort and often fail to meet the performance requirements of data-intensive applications running at large scale.

In this work we go through the basic requirements of storing telemetry data streams for computational steering and visualization. For simple use cases where the telemetry data is only temporary, we prove that distributed logging can be performed at scale by leveraging state-of-the-art blob storage systems such as Týr or RADOS. This approach is supported by the growing availability of node-local storage on a new generation of supercomputers, giving application developers the freedom to deploy transient storage systems alongside the application directly on the compute nodes.

When long-term storage of the generated data is needed for offline visualization or analytics, we prove that distributed logs require a significantly lower number of output logs to achieve peak performance compared to Lustre or GPFS. We also prove that this low number of output files obviates the need for an explicit post-processing merge step in most cases for iterating the whole output log in generation order. We finally prove on up to 100,000 cores of the Theta supercomputer that our findings are applicable to run distributed logging at large scale, while improving write throughput by several orders of magnitude compared to Lustre of GPFS.

6.3.2. Leveraging Damaris for in-situ visualization in support of GeoScience and CFD Simulations

Participants: Hadi Salimi, Matthieu Dorier, Luc Bougé.

Damaris is a middleware for in situ data analysis and visualization targeting extreme-scale, MPI-based simulations. The main goal of Damaris is to provide a simple method to instrument a simulation in order to benefit from in situ analysis and visualization. To this aim, the computing resources are partitioned such that a subset of cores in a SMP node or a subset of nodes of the underlying platform are dedicated to in situ processing. The data generated by the simulation are passed to these dedicated processes either through shared memory (in the case of dedicated cores) or through the MPI calls (in the case of dedicated nodes) and can be processed both in synchronous and asynchronous modes. Afterwards, the processed data can be analyzed or visualized. Damaris also supports a very simple API to instrument simulations developed in different domains. Moreover, using some XML configuration files for defining simulation data types (e.g. meshes) makes the instrumentation process easier with minimum code modifications. Active development is currently continuing within the KerData team, where it is at the center of several collaborations with industry (e.g Total) as well as with national and international academic partners.

In recent developments of Damaris, we have focused on two main targets that are: 1) Instrumenting new simulations codes from different scientific domains, i.e. geoscience and ocean modeling, 2) Implementing new storage backends, i.e. HDF5 for Damaris. In this regard, we report the results of some experiments we made to evaluate Damaris with respect to performance. These experiments were conducted on Grid'5000 test bed. In these experiments Damaris was employed to visualize the data generated by the Wave Propagation geoscience simulation and also the CROCO coastal and ocean simulation. During the experiments, the impact of Damaris was measured by comparing the simulations instrumented by Damaris (space partitioning approach) with a baseline where those simulations include data processing codes directly on their source code (time partitioning approach). The results of these simulations show that the incorporation of Damaris into a simulation decreases the total run time of the simulation due to its asynchronous data processing and visualization capabilities. In addition, using Damaris for data visualization has nearly no impact on the total run time of the mentioned simulation codes. We also have shown that the amount of code changes necessary for instrumenting the simulation codes is much less compared to the case that the simulation code is instrumented by native visualization or storage APIs. Moreover, we also have studied the impact of new HDF5 storage backend, on storing simulation results in HDF5 format in both file-per-dedicated-core and collective I/O scenarios.

6.3.3. Accelerating MPI collective operations on the Theta cupercomputer

Participants: Nathanaël Cheriere, Matthieu Dorier, Misbah Mubarak, Robert Ross, Gabriel Antoniu.

Recent network topologies in supercomputers have motivated new research on topology-aware collective communication algorithms for MPI. But such endeavor requires betting on the fact that topology-awareness is the primary factor to accelerate these collective operations. Besides, showing the benefit of a new, topology-aware algorithm requires not only access to a leadership-scale supercomputer with the desired topology, but also a large resource allocation on this supercomputer. Event-driven network simulations can alleviate such constraints and speed up the search for appropriate algorithms by providing early answers on their relative merit.

In our studies, we focus on the Scatter and AllGather operations in the context of the Theta supercomputer's dragonfly topology. We propose a set of topology-aware versions of these operations as well as optimizations of the old, non-topology-aware ones. We conduct an extensive simulation campaign using the CODES network simulator. Our results show that, contrary to our expectations, topology-awareness does not help improving significantly the speed of these operations. Rather, the high radix and low diameter of the dragonfly topology, along with already good routing protocols, enable simple algorithms based on non-blocking communications to perform better than state-of-the-art algorithms. A trivial implementation of Scatter using nonblocking point-to-point communications can be faster than state-of-the art algorithms by up to a factor of 6. Traditional AllGather algorithms can also be improved by the same principle and exhibit a 4x speedup in some

situations. These results highlight the need to rethink the collective operations under the light of nonblocking communications.

6.4. Energy-aware data storage and processing at large scale

6.4.1. Performance and energy-efficiency trade-offs in in-memory storage systems

Participants: Mohammed-Yacine Taleb, Shadi Ibrahim, Gabriel Antoniu, Toni Cortes.

Most large popular web applications, like Facebook and Twitter, have been relying on large amounts of in-memory storage to cache data and offer a low response time. As the main memory capacity of clusters and clouds increases, it becomes possible to keep most of the data in the main memory. This motivates the introduction of in-memory storage systems. While prior work has focused on how to exploit the low-latency of in-memory access at scale, there is very little visibility into the energy-efficiency of in-memory storage systems. Even though it is known that main memory is a fundamental energy bottleneck in computing systems (i.e., DRAM consumes up to 40% of a server's power). During this project, by the means of experimental evaluation, we have studied the performance and energy-efficiency of RAMCloud - a well-known in-memory storage system. We reveal that although RAMCloud is scalable for read-only applications, it exhibits non-proportional power consumption. We also find that the current replication scheme implemented in RAMCloud limits the performance and results in high energy consumption. Surprisingly, we show that replication can also play a negative role in crash-recovery.

6.4.2. Energy-aware straggler mitigation in Map-Reduce

Participants: Tien-Dat Phan, Chi Zhou, Shadi Ibrahim, Guillaume Aupy, Gabriel Antoniu.

Energy consumption is an important concern for large-scale data-centers, which results in huge monetary cost for data-center operators. Due to the hardware heterogeneity and contentions between concurrent workloads, straggler mitigation is important to many Big Data applications running in large-scale data-centers and the speculative execution technique is widely-used to handle stragglers. Although a large number of studies have been proposed to improve the performance of Big Data applications using speculative execution, few of them have studied the energy efficiency of their solutions.

In [23], we propose two techniques to improve the energy efficiency of speculative executions while ensuring comparable performance. Specifically, we propose a hierarchical straggler detection mechanism which can greatly reduce the number of killed speculative copies and hence save the energy consumption. We also propose an energy-aware speculative copy allocation method which considers the trade-off between performance and energy when allocating speculative copies. We implement both techniques into Hadoop and evaluate them using representative Map-Reduce benchmarks. Results show that our solution can reduce the energy waste on killed speculative copies by up to 100% and improve the energy efficiency by 20% compared to state-of-the-art mechanisms.

MYRIADS Project-Team

7. New Results

7.1. Scaling Clouds

7.1.1. Fog Computing

Participants: Guillaume Pierre, Arif Ahmed, Ali Fahs, Alexandre Van Kempen, Salsabil Amri, Vinothkumar Nagasayanan, Berenger Nguyen Nhon.

Fog computing aims to extend datacenter-based cloud platforms with additional compute, networking and storage resources located in the immediate vicinity of the end users. By bringing computation where the input data was produced and the resulting output data will be consumed, fog computing is expected to support new types of applications which either require very low network latency (e.g., augmented reality applications) or which produce large data volumes which are relevant only locally (e.g., IoT-based data analytics).

Fog computing architectures are fundamentally different from those of traditional cloud platforms: to provide computing resources in physical proximity of any end user, fog computing platforms must necessarily rely on very large numbers of small Points-of-Presence connected to each other with commodity networks whereas clouds are typically organized with a handful of extremely powerful data centers connected by dedicated ultra-high-speed networks. This geographical spread also implies that the machines used in any Point-of-Presence may not be datacenter-grade servers but much weaker commodity machines.

We investigated the challenges of efficiently deploying Docker containers in fog platforms composed of tiny single-board computers such as Raspberry PIs. This operation can be painfully slow, in the order of multiple minutes depending on the container's image size and network condition. We showed that this bad performance is not only due to hardware limitations, but it is largely due to inefficiencies in the way Docker implements the container's image download operation. We proposed a number of optimization techniques which , when combined together, make container deployment up to 4 times faster than the vanilla Docker implementation. A publication on this topic is under submission.

Although fog computing infrastructures are fundamentally distributed, their management part still remains centralized: a single node (or small group of nodes) is in charge of maintaining the list of available server machines, monitoring them, distributing software to them, deciding which server must take care of which task, etc. We therefore aim to reduce the discrepancy between the broadly distributed compute/storage resources and the – currently – extremely centralized control of these resources, by focusing first on the resource scheduling function. This project has just started, and we expect to obtain the first results in 2018.

7.1.2. Edge Cloud

Participants: Anne-Cécile Orgerie, Cédric Tedeschi, Matthieu Simonin, Ehsan Ahvar, Genc Tato.

Myriads is involved in the Discovery project, whose goal is to design, develop and experiment a software stack for a distributed cloud platform where resources are directly injected into the backbone of the network [60]. To this end, we designed a novel family of overlay network to operate messaging and routing on top of such a distributed utility computing platform. The big picture of these overlays was described in a workshop [47].

7.1.3. Community Clouds

Participant: Jean-Louis Pazat.

In this work we consider an infrastructure based on devices (such as Internet boxes and NAS) owned and operated by end-users. A typical use-case is the sharing of CPU and storage capabilities by a community of users. This sharing is operated by hosting services to local and remote users. The devices of this distributed infrastructure have heterogeneous capabilities and no guaranteed availability. It is therefore challenging to ensure to the guest service a minimal hosting service level, such as availability or QoS.

We consider services build as an application based on micro-services. Such an application is deployed on the infrastructure by instantiating its constituant micro-services on some devices. One micro-services may rely on others micro-services to enable its own service. The performance of the resulting application is therefore highly dependent from the placement for each micro-service instance. Device parameters like CPU capabilities or network bandwidth and latency have a significant impact on the resulting response time of the micro-service, hence the application.

We explore solutions to adapt the placement of the micro-services to the capabilities of the infrastructure. As a first step, we are studying a static system where these capabilities are not variating. The placement decision can be expressed as the solution of an NP-Complete optimization problem. We have shown that a solution for this problem can be found with reasonably good precision using a meta-heuristic called Particle Swarm Optimization. The next step will be to study how this solution can be adapted in a dynamic system by considering the variations of the CPU and Network parameters and the availability of the devices.

This work in done in the context of Bruno Stevant's PhD thesis co-advised by Jean-Louis Pazat (Bruno Stevant is a member of OCIF team).

7.1.4. Evaluation of Data Stream Processing Frameworks in Clouds

Participants: Christine Morin, Deborah Agarwal, Subarna Chatterjee.

We address the problem of selecting a correct stream processing framework for a given application to be executed within a specific physical infrastructure. For this purpose, we have performed a thorough comparative analysis of three data stream processing platforms – Apache Flink, Apache Storm, and Twitter Heron (the enhanced version of Apache Storm), that are chosen based on their potential to process both streams and batches in real-time. For the comparative performance analysis of the chosen platforms, we have experimented using 8-node clusters on Grid5000 experimentation testbed and have selected a wide variety of applications ranging from a conventional benchmark (word count application) to sensor-based IoT application (air quality monitoring application) and statistical batch processing application (flight delay analysis application). The work focuses to analyze the performance of the frameworks in terms of the volume and throughput of data streams that each framework can possibly handle. The impact of each framework on the operating system is analyzed by experimenting and studying the resource utilization of the platforms in terms of CPU utilization, memory consumption. The energy consumption of the platforms is also studied to understand the suitability of the platforms towards green computing. Last, but not the least, the fault tolerance of the frameworks is also studied and analyzed. Lessons learnt from this work will precisely enlighten IaaS cloud end-users to wisely choose the correct streaming platform in order to run a particular application within a given set of VMs and will assist the cloud-providers to rationally allocate VMs equipped with a particular stream processing framework to PaaS cloud-users for running a specific streaming application. A paper has been submitted to an international conference in November 2017.

7.1.5. Stream Processing for Maritime Surveillance

Participants: Pascal Morillon, Christine Morin, Matthieu Simonin, Cédric Tedeschi.

In the context of maritime surveillance, and of the Sesame Project, we started the design and implementation of a platform dedicated to the batch and real-time processing of AIS messages sent by ships to inform about their identity, position and destination among other pieces of information.

Having use cases in mind such as detecting ships entering a protected areas, or ships having suspect behaviors, we designed a software architecture able to process AIS messages and produce synthetic data so as to answer these questions.

First experiments using a preliminary version of this platform have been conducted over the Grid'5000 platform using an archive of one-month of the AIS messages collected globally during March 2017. In particular, we've been able to index these messages using ElasticSearch [0] and visualize them using Kibana [0].

[0] https://www.elastic.co/fr/
[0] https://www.elastic.co/products/kibana

The architecture has been described in a poster presented at BiDS'17 [56].

7.1.6. Adaptive deployment for multi-cloud applications

Participants: Nikos Parlavantzas, Manh Linh Pham.

This work builds on the Adapter system, developed in the context of the PaaSage European project (2012-2016). The Adapter is part of the PaaSage open-source platform, a holistic solution for supporting the automatic deployment and execution of multi-cloud applications. Specifically, the Adapter is responsible for dynamic, cross-cloud application adaptation, taking into account adaptation costs and benefits in making deployment decisions. In 2017, we improved the Adapter and performed a comprehensive evaluation using experiments in a multi-cloud environment. The results demonstrate that Adapter supports automated multi-cloud adaptation while optimizing the performance and cost of the application. The results are described in an article currently under submission.

7.1.7. Application configuration and reconfiguration in multi-cloud environments

Participant: Nikos Parlavantzas.

Current approaches to cloud application configuration and reconfiguration are typically platform dependent, error prone and provide little support for optimizing application performance and resource utilisation. To address these limitations, we are combining the use of software product lines (SPLs) with performance prediction and automatic adaptation techniques. This work is performed in the context of the thesis of Carlos Ruiz Diaz, a PhD student at the University of Guadalajara, co-advised by Nikos Parlavantzas. The work has produced an SPL-based framework supporting initial configuration and dynamic adaptation in a systematic, platform-independent way.

In 2017, we extended the framework with a proactive adaptation solution that performs vertical VM scaling based on predictions of resource utilisation and performance. The solution targets multi-tier applications deployed on IaaS clouds. Experimental results demonstrate that the solution maintains expected application performance while reducing resource waste [46].

7.1.8. Adaptive resource management for high-performance, multi-sensor systems

Participants: Christine Morin, Nikos Parlavantzas, Baptiste Goupille–Lescar.

In the context of our collaboration with Thales Research and Technology and Baptiste Goupille-Lescar's PhD work, we are applying cloud resource management techniques to high-performance, multi-sensor, embedded systems with real-time constraints. The objective is to increase the flexibility and efficiency of resource allocation in such systems, enabling the execution of dynamic sets of applications with strict QoS requirements.

In 2017, we focused on an industrial use case concerning the operation of a multi-function surface active electronically scanned array (AESA) radar. We developed a simulation environment using an industrial high-precision AESA simulator and the Ptolemy II simulation framework, and we are using this environment to explore and evaluate different dynamic application placement solutions [57].

7.2. Greening Clouds

ICT (Information and Communications Technologies) ecosystem now approaches 6% of world electricity consumption and this ICT energy use will continue grow fast because of the information appetite of Big Data, big networks and big infrastructures as Clouds that unavoidably leads to big power.

7.2.1. Energy Models

Participants: Ehsan Ahvar, Loic Guegan, Anne-Cécile Orgerie, Martin Quinson.

Cloud computing allows users to outsource the computer resources required for their applications instead of using a local installation. It offers on-demand access to the resources through the Internet with a pay-as-you-go pricing model.However, this model hides the electricity cost of running these infrastructures.

The costs of current data centers are mostly driven by their energy consumption (specifically by the air conditioning, computing and networking infrastructure). Yet, current pricing models are usually static and rarely consider the facilities' energy consumption per user. The challenge is to provide a fair and predictable model to attribute the overall energy costs per virtual machine and to increase energy-awareness of users. We aim at proposing such energy cost models without heavily relying on physical wattmeters that may be costly to install and operate.

Another goal consists in better understanding the energy consumption of computing and networking resources of Clouds in order to provide energy cost models for the entire infrastructure including incentivizing cost models for both Cloud providers and energy suppliers. These models will be based on experimental measurement campaigns on heterogeneous devices. Inferring a cost model from energy measurements is an arduous task since simple models are not convincing, as shown in our previous work. We aim at proposing and validating energy cost models for the heterogeneous Cloud infrastructures in one hand, and the energy distribution grid on the other hand. These models will be integrated into simulation frameworks in order to validate our energy-efficient algorithms at larger scale.

7.2.2. Exploiting Renewable Energy in Clouds

Participants: Benjamin Camus, Yunbo Li, Anne-Cécile Orgerie.

The development of IoT (Internet of Things) equipment, the popularization of mobile devices, and emerging wearable devices bring new opportunities for context-aware applications in cloud computing environments. The disruptive potential impact of IoT relies on its pervasiveness: it should constitute an integrated heterogeneous system connecting an unprecedented number of physical objects to the Internet. Among the many challenges raised by IoT, one is currently getting particular attention: making computing resources easily accessible from the connected objects to process the huge amount of data streaming out of them.

While computation offloading to edge cloud infrastructures can be beneficial from a Quality of Service (QoS) point of view, from an energy perspective, it is relying on less energy-efficient resources than centralized Cloud data centers. On the other hand, with the increasing number of applications moving on to the cloud, it may become untenable to meet the increasing energy demand which is already reaching worrying levels. Edge nodes could help to alleviate slightly this energy consumption as they could offload data centers from their overwhelming power load and reduce data movement and network traffic. In particular, as edge cloud infrastructures are smaller in size than centralized data center, they can make a better use of renewable energy.

We propose to investigate the end-to-end energy consumption of IoT platforms. Our aim is to evaluate, on concrete use-cases, the benefits of edge computing platforms for IoT regarding energy consumption. We aim at proposing end-to-end energy models for estimating the consumption when offloading computation from the objects to the edge or to the core Cloud, depending on the number of devices and the desired application QoS, in particular trading-off between performance (response time) and reliability (service accuracy).

7.2.3. Smart Grids

Participants: Benjamin Camus, Anne-Cécile Orgerie, Martin Quinson.

We propose exploiting Smart Grid technologies to come to the rescue of energy-hungry Clouds. Unlike in traditional electrical distribution networks, where power can only be moved and scheduled in very limited ways, Smart Grids dynamically and effectively adapt supply to demand and limit electricity losses (currently 10% of produced energy is lost during transmission and distribution).

For instance, when a user submits a Cloud request (such as a Google search for instance), it is routed to a data center that processes it, computes the answer and sends it back to the user. Google owns several data centers spread across the world and for performance reasons, the center answering the user's request is more likely to be the one closest to the user. However, this data center may be less energy efficient. This request may have consumed less energy, or a different kind of energy (renewable or not), if it had been sent to this further data center. In this case, the response time would have been increased but maybe not noticeably: a different trade-off between quality of service (QoS) and energy-efficiency could have been adopted.

While Clouds come naturally to the rescue of Smart Grids for dealing with this big data issue, little attention has been paid to the benefits that Smart Grids could bring to distributed Clouds. To our knowledge, no previous work has exploited the Smart Grids potential to obtain and control the energy consumption of entire Cloud infrastructures from underlying facilities such as air conditioning equipment (which accounts for 30% to 50% of a data center's electricity bill) to network resources (which are often operated by several actors) and to computing resources (with their heterogeneity and distribution across multiple data centers). We aim at taking advantage of the opportunity brought by the Smart Grids to exploit renewable energy availability and to optimize energy management in distributed Clouds.

7.2.4. Involving Users in Energy Saving

Participants: David Guyon, Christine Morin, Anne-Cécile Orgerie.

In a Cloud moderately loaded, some servers may be turned off when not used for energy saving purpose. Cloud providers can apply resource management strategies to favor idle servers. Some of the existing solutions propose mechanisms to optimize VM scheduling in the Cloud. A common solution is to consolidate the mapping of the VMs in the Cloud by grouping them in a fewer number of servers. The unused servers can then be turned off in order to lower the global electricity consumption.

Indeed, current work focuses on possible levers at the virtual machine suppliers and/or services. However, users are not involved in the choice of using these levers while significant energy savings could be achieved with their help. For example, they might agree to delay slightly the calculation of the response to their applications on the Cloud or accept that it is supported by a remote data center, to save energy or wait for the availability of renewable energy. The VMs are black boxes from the Cloud provider point of view. So, the user is the only one to know the applications running on her VMs.

We plan to explore possible collaborations between virtual machine suppliers, service providers and users of Clouds in order to provide users with ways of participating in the reduction of the Clouds energy consumption. This work will follow two directions: 1) to investigate compromises between power and performance/service quality that cloud providers can offer to their users and to propose them a variety of options adapted to their workload; and 2) to develop mechanisms for each layer of the Cloud software stack to provide users with a quantification of the energy consumed by each of their options as an incentive to become greener.

Our results were published in [40], [32], [31].

7.3. Securing Clouds

7.3.1. Security Monitoring in Clouds

Participants: Christine Morin, Jean-Louis Pazat, Louis Rilling, Anna Giannakou, Amir Teshome Wonjiga, Clément El Baz.

In the INDIC project 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 verifying SLA terms,
- making the SLA terms enforcement mechanisms self-adaptable to cope with the dynamic nature of clouds.

The considered enforcement and verification mechanisms should have a minimal impact on performance.

In 2017, we did a thorough performance evaluation and security correctness analysis of the SAIDS approach, that we proposed in 2015, and that makes a network intrusion detection system (NIDS) deployed in a cloud operator infrastructure self-adaptable. In the performance evaluation we studied the performance impact of SAIDS on the cloud infrastructure operations related to the management of virtual machines (typically creation, migration, and deletion) as well as the scalability of SAIDS with respect to the number of NIDS devices managed. This performance evaluation was done on the Grid'5000 platform. The results showed that SAIDS adds very low overhead and is scalable. The security analysis was done both experimentally and based on a risk analysis. This analysis validated the security correctness of SAIDS. A full paper presenting SAIDS and its evaluation is submitted for publication in 2018. A demo of SAIDS was presented at FIC 2017, Lille, France in January 2017 and at the Inria Industry Days, Paris, France on October 17th, 2017.

Regarding SLA definition and enforcement, in 2017 we evaluated the verification method that we defined in 2016 and that enables a Cloud customer to verify that an NIDS located in the operator infrastructure is configured correctly according to the Service-Level Objectives (SLO) figuring in the SLA. The performance evaluation was done on the Grid'5000 platform and showed that the proposed verification method requires making a trade-off between verification speed and impact on the performance of the production applications deployed in the tenant's virtual machines. The security correctness analysis was based on a risk analysis and showed the constraints on the types of attacks that can be used for verification as well as the limitations due to the tools used in the prototype [55]. A full paper presenting the verification method and its evaluation is submitted for publication in 2018.

After the acquired experience on verifying security monitoring metrics, we started studying how to define relevant SLOs that are verifiable. We plan to get results in 2018 and submit a paper for publication in 2018 or 2019.

Finally, in October 2017 we started studying how security monitoring SLAs could take into account context changes like the evolution of threats and updates to the tenants' software.

Our work done as part of the INDIC project were presented in [59].

7.3.2. Risk assessment in clouds

Participant: Christine Morin.

Cloud providers have an incomplete view of their hosted virtual infrastructures managed by a Cloud Management System (CMS) and a Software Defined Network (SDN) controller. For various security reasons (e.g. isolation verification, modeling attack paths in the network), it is necessary to know which virtual machines can interact via network protocols. This requires building a connectivity graph between the virtual machines, that we can extract with the knowledge of the overall topology and the deployed network security policy. Existing methodologies for building such models for physical networks produce incomplete results. Moreover, they are not suitable for cloud infrastructures due to either their intrusiveness or lack of connectivity discovery. We propose a method to compute the connectivity graph, relying on information provided by both the CMS and the SDN controller. Connectivity can first be extracted from knowledge databases, then dynamically updated on the occurrence of cloud-related events. We realized an experimental evaluation of the proposed method to determine its correctness and performance in a realistic context, considering CPU and RAM consumption, the volume of data generated, and execution time for the different portions of the algorithm involved. Experiments were run on the Grid'5000 platform with OpenStack CMS and ONOS SDN controller. Our approach proves on a representative infrastructure to compute exact, complete and up-to- date connectivity graphs in reasonable time [42], [41].

7.3.3. Personal Data Management in Cloud-based IoT Systems

Participants: Christine Morin, Jean-Pierre Banâtre, Deborah Agarwal, Subhadeep Sarkar, Louis Rilling.

The Internet of Things (IoT), in today's digital world, encompasses billions of smart connected devices. These devices generate an unprecedented amount of data, which often bears sensitive personal information of individuals. In present service models, the data are processed and managed by service providers, beyond the visibility of the owner of the data. Although the EU General Data Protection Regulation (GDPR) strives to protect citizens and their data by regulation, citizens and service providers need technological advances to gain effective control over their data or to prove compliance with the new regulation. Our primary objective is to enforce, by design, the GDPR at the system level so as to preserve the privacy concerning personal data. We started off with enforcement of the data erasure facility as expressed in the GDPR. Data erasure corresponds to both automatic erasure of data after expiration of their retention period and ad-hoc on request of the data owner. Our first contribution, towards this, is design of a customizable privacy policy, which would allow the end users to express their preferences regarding the purpose of use, location of processing, retention period, sharing and storage policies concerning their personal data. We developed a XML-based policy expression language by defining the required data structures and vocabulary, which will facilitate the end-users to easily express their preferences. Next, we have investigated into the possible way of the implementation of the proposed solution and identified the exploitation of the operation system capabilities as an appropriate means to the cause. For this, we have potentially chosen the Sel4 (or may be some other capability-based microkernel) as our platform of operation. Finally, we have identified the different challenges towards implementation of our solution and did some groundwork towards proposing the solutions to the same. These challenges include efficient identification of replication of data, locating all replicas of a given data segment, and implementing erasure of data in a cross-domain service model.

7.4. Experimenting with Clouds

7.4.1. Simulation

Participants: Martin Quinson, Loic Guegan, Toufik Boubehziz, The Anh Pham.

We propose to combine two complementary experimental approaches: direct execution on testbeds such as Grid'5000, that are eminently believable but rather labor intensive, and simulations (using *e.g.* SimGrid) that are much more light-weighted, but requires are careful assessment. One specificity of the Myriads team is that we are working on these experimental methodologies *per se*, raising the standards of *good experiments* in our community. The Grid'5000 operational team is embedded in our research team, ensuring that our work remains aligned with the ground reality.

In 2017, our work was mostly centered on letting SimGrid become a *de facto* standard for the simulation of distributed platforms. We introduced a new programming interface, particularly adapted to the study of abstract algorithms. Beyond the engineering task, this requires to carefully capture the concepts that are important to the practitioners on distributed systems.

SimGrid is not limited to abstract algorithms, and can also be used to simulate real applications. This year, we published a journal article on the many challenges to overcome when designing a simulator of high performance systems. This work was published in the TPDS journal [20].

On the modeling side, our team worked this year toward the improvement of energy models, both for computational facilities and for the network. Despite the scarce availability of real testbeds that allow fine-grained energy measurements, we managed to provide a generic energy consumption model, published in [35], [43].

Finally, we restarted our efforts toward the formal verification of distributed systems. The model-checker that is integrated within SimGrid is already functional ([44]), but more work is necessary to make it efficient. We even found cases for which our reduction algorithm may miss defects in the verified system. This work will certainly motivate much more work in the future years.

7.4.2. Use cases

Participants: Christine Morin, Nikos Parlavantzas, Deborah Agarwal, Manh Linh Pham.

7.4.2.1. Simulation framework for studying between-herd pathogen spread in a region

In the context of the MIHMES project (2012-2017) and in collaboration with INRA researchers, we transformed a legacy application for simulating the spread of bovine viral diarrhea virus (BVDV) to a cloud-enabled application based on the DiFFuSE framework (Distributed framework for cloud-based epidemic simulations). Specifically, the original sequential code was first modified to add single-computer parallelism using OpenMP. We then decomposed the code into separate services that were deployed across multiple clouds and independently scaled. Using this service-based cloud-enabled simulation, we performed a set of experiments that demonstrated that applying DiFFuSE increases performance, allows exploring different cost-performance trade-offs, automatically handles failures, and supports elastic allocation of resources from multiple clouds [45].

7.4.2.2. FluxNet and AmeriFlux Data Analysis

The carbon flux datasets from AmeriFlux (Americas) and FLUXNET (global) are comprised of long-term time series data and other measurements at each tower site. There are over 800 flux towers around the world collecting this data. The non-time series measurements include information critical to performing analysis on the site's data. Examples include: canopy height, species distribution, soil properties, leaf area, instrument heights, etc. These measurements are reported as a variable group where the value plus information such as method of measurement and other information are reported together. Each variable group has a different number and type of parameters that are reported. The current output format is a normalized file. Users have found this file difficult to use.

Our earlier work in the DALHIS Inria associate team focused on building user interfaces to specify the data. This year we jointly worked on developing a Jupyter Notebook that would serve as a tool for users to read in and explore the data in a personalized tutorial type environment. We developed two notebooks and the next step is to start user testing on the notebooks.

TACOMA Team

6. New Results

6.1. Smart City and ITS

Participants: Indra Ngurah, Djibrilla Amadou Kountche, Xavier Gilles, Christophe Couturier, Rodrigo Silva, Frédéric Weis, Jean-Marie Bonnin [contact].

The domain of Smart Cities is still young but it is already a huge market which attract number of companies and researchers. It is also multi-fold as the words "smart city" gather multiple meanings. Among them one of the main responsibilities of a city, is to organisation the transportation of goods and people. In intelligent transportation systems (ITS), ICT technologies have been involved to improve planification and more generally efficiency of journeys within the city. We are interested in the next step where efficiency would be improved locally relying on local interactions between vehicles, infrastructure and people (smartphones).

For the future "autonomous" vehicle are now in the spotlight, since a lot of works has been done in recent years in automotive industry as well as in academic research centers. Such unmanned vehicle could strongly impact the organisation of the transportation in our cities. However, due to the lack of a definition of what is an "autonomous" vehicle it remains still difficult to see how these vehicles will interact with their environment (eg. road, smart city, houses, grid, etc"). From augmented perception to fully cooperative automated vehicle, the autonomie cover various realities in terms of interaction the vehicle relies on. The extended perception relies on communication between the vehicle and surrounding roadside equipments. This help the driving system to build and maintain an accurate view of the environment. But at this first stage the vehicle only uses its own perception to make its decisions. At a second stage, it will take benefits of local interaction with other vehicles through car-to-car communications to elaborate a better view of its environment. Such "cooperative autonomy" does not try to reproduce the human behavior anymore, it strongly rely on communication between vehicles and/or with the infrastructure to make decision and to acquire information on the environment. Part of the decision could be centralized (almost everything for an automatic metro) or coordinated by a roadside component. The decision making could even be fully distributed but this put high constraints on the communications. Automated vehicles are just an exemple of smart city automated processes that will have to share information within the surrounding to make their decisions.

We participated in the definition of the distributed architecture that has been adopted by all partners of the SEAS project. The main principles of this architecture have been published and we developed several profs of concept that have been demonstrated in the project consortium. Our partner developed the components of the architecture that has been demonstrated in the final review of the project (in January). The principles of the architecture and data representation has been used to design an open reusable Data Manager in the context of the EkoHub projet. This modular software will be extended to fit the needs of Indra Ngurah and Rodrigo Silva works.

6.2. Convergence middleware for pervasive data

Participants: Yoann Maurel, Jules Desjardin, Paul Couderc [contact].

We are currently working on data driven middleware approaches dedicated to physical objects and smart spaces. We had previous contributions on the topic, where opportunistic collaborations between mobile devices were supported by Linda-like tuple space and IEEE 802.11 radios. However, these were adapted to relatively complex devices and the technological limitation at the time did not allow the full potential of the model. More recently, we investigated distributed storage spread over physical objects or fragments using RFID, enabling complex data to be directly reflected by passive objects (without energy). Yet other radio technologies, such as BLE, are emerging to support close range interactions with very low (or even zero) energy requirements.

Applications such as pervasive games (for ex. Pokemon Go), on the go data sharing, collaborative mobile app are often good candidates for opportunistic or dynamic interaction models. But they are not well supported by existing communication stacks, especially in context involving multiple technologies. Technological heterogeneity is not hidden, and high level properties associated with the interactions, such as proximity/range, or mobility-related parameters (speed, discovery latency) have to be addressed in an ad hoc manner. We think that a good way to solve these issues is to offer an abstract interaction model that could be mapped over the common proximity communication technologies, in a similar way as MOM (Message Oriented Middleware) such as MQTT abstract communications in many IoT and pervasive computing scenarios. However, they typically requires IP level communication, which far beyond the capabilities of ultra low energy proximity communication such as RFID and BLE. Moreover, they often rely on a coordinator node that is not adapted in highly dynamic context involving ephemeral communications and mobile nodes.

We started the implementation of an associative memory mechanism over BLE, as it is a common ground that can be shared with passive or semi passive communications (RFID, NFC). Such mechanism, although relatively low level, is still a very useful building block for opportunistic applications: it enables opportunistic data storage/sharing and signaling/synchronization (in space in particular). This approach is fully in line with more general trend developed in the team to build "smart" systems leveraging local resources and data oriented mediation. We have started validation work with a few applications, in particular regarding energy aspects and scalability with respect to the communication load. This should lead to publishing on both infrastructure and application level aspects of the approach.

6.3. Modeling activities and forecasting energy consumption and production to promote the use of self-produced electricity from renewable sources

Participants: Alexandre Rio, Yoann Maurel [contact].

This work began in 2017 and is carried out as part of a broader collaboration between Tacoma, the Diverse Team and OKWind, a company specialized in the production of renewable sources of energy. OKWind proposes to deploy self-production units directly where the consumption is done. It has developed expertise in vertical-axis wind turbines, photovoltaic trackers, and heat pump. This project aims at building a system that optimizes the use of different sources of renewable energy, choosing the most suitable source for the current demand and anticipating future needs. The goal is to favor the consumption of locally produced electricity and to maximize the autonomy of the equipped sites so as to reduce the infrastructure needed to distribute electricity, to set energy cost, and to reduce the ecological impact of energy consumption.

Modeling and forecasting production and consumption of a site is hard and raises several issues: how to precisely assess the consumption and production of energy on a given site with changing conditions ? How to adequately size energy sources and energy storage (wind turbine, solar panel and batteries) ? And what methods to use to optimize consumption and, whenever possible, act on installations and activities to reduce energy costs ? We aim to propose tools to predict the consumption of a site based on estimation and previous observation, monitor the site in real time and forecast evolution. We propose to build a DSL describing consumption and production processes, and a system providing recommendations based on the derived model at runtime.

The problem of forecasting is known from both a production and consumption point of view. OKWind has developed tools to predict the production of their renewable sources - the same goes for batteries - and a lot of theoretical work has been done on consumption in the literature. In our view little has been done to precisely model activities, their energy consumption and the associated variability. Indeed most of the current approaches are concerned with either large-scale forecasting for the Smart Grid, are based on coarse grain data (total energy consumption of the site), or focuses on modeling specific appliance without describing how and when they are used.

This is paradoxical considering that companies have spent a lot of time modeling their activities from a logistic point of view. Intuitively, the predictable and seasonal nature of a company's activities would allow building activity schedulers that favor the consumption of certain energy sources (the cleanest or least expensive one for

instance). The development of a DSL to describe the relationships between activities, their planning, and the production and environmental factors would make possible to simulate a given site at a given location, to make assumptions on sizing, and would be a basis to forecast energy consumption so as to provide recommendations for the organization of activities.

We already have developed part of this DSL that simulates activities and production. In particular, it is capable of simulating consumption and production over a given period based on available environmental data. This tool is in the experimentation phase. In particular, we are collecting information on several sites to measure the consumption of various activities.

6.4. Sharing knowledge and access-control

Participants: Adrien Capaine, Yasmina Andaloussi, Frédéric Weis, Yoann Maurel [contact].

Smart spaces (Smart-city, home, building, etc.) are complex environments made up of resources (cars, smartphones, electronic equipment, applications, servers, flows, etc.) that cooperate to provide a wide range of services to a wide range of users. They are by nature extremely fluctuating, heterogeneous, and unpredictable. In addition, applications are often mobile and have to migrate or are offered by mobile platforms such as smartphones or vehicles.

To be relevant, applications must be able to adapt to users by understanding their environment and anticipating its evolutions. They are therefore based, explicitly or implicitly, on a representation of their surrounding environment based on available data provided by sensors, humans, objects and applications when available. The accuracy of the adaptations made by the applications depends on the precision of this representation. Building and maintaining such knowledge is resource-intensive in terms of network exchanges, computing time and incidentally energy consumption. It is, therefore, crucial to find ways to improve this process. In practice, many applications build their own models without sharing them or delegating calculations to remote services, which is not optimal because many processes are redundant. A huge improvement would be to find mechanisms that allows sharing the information so as to reduce as much as possible the treatments necessary to obtain it.

However, it seems extremely complex to provide a global, complete and unified view of the environment that reflects the applications' concerns. If it were possible, such a single representation would by nature be incomplete or subjective. Our solution should be applicable to nowadays devices and applications with little adjustments to the underlying architectures. It should then be flexible enough to deal with the lack of standards in the domain without imposing architectural choices. Such lack of standard is very common in IT and mainly due to well-known factors: (1) for technical reasons, developers tend to think that their "standard" is better suited for their current use-case, or/and (2) for commercial reasons companies want to keep a closed siloed system to capture their users, or/and (3) because the domain is still new and evolving and no standard as emerged yet, or/and finally (4) because the problem is too complex to be standardized and most proposed standards tend to be bloated and hard to use. The IoT domain suffers from all of these impediments and solution targeting mid-term application have to take these factors into accounts. Many IoT applications are still organized in silos of information. This leads to the deployment of sensors with similar functions and redundant pieces of software providing exactly the same service. Many frameworks or ontologies have been developed in the field to provide a solution to this problem but their implementation depends on the goodwill of the companies who do not always see their interest in losing part of the control of their application and data. To be largely accepted, solutions should let companies decide what information to share and when with little impact on their current infrastructure.

We want to be able to develop collaborative mechanisms that allow applications to share some of their information about the immediate surrounding environment with their counterparts. The idea is to allow the construction of shared representations between groups of applications that manipulate the same concepts so that each group can construct a subjective and complete representation of the environment that corresponds to its concerns. In this context, we want to offer applications mechanisms allowing them to leave information about their environment by associating them directly with the flows, data, services and objects handled. This

information will be stored by the environment so that it will be possible for the application to retrieve it and for its peers to access it. From a logical point of view, applications will have the illusion of annotating objects directly; we make no assumptions about where this information will be stored, which will depend on the characteristics of the environment or the sharing solution chosen. Data should be stored as close as possible to the environments they qualify for reasons of performance, confidentiality and autonomy. To experience that idea, we have developed:

- Matriona, a globally distributed framework developed on top of OSGi. This project has been described in more details in the previous activity report. It is meant to be a global framework for exposing devices as REST-like resources. Resources functionalities can be extended through the mean of decorators. The system also provides access mechanisms. The main interest of Matriona with regards to the information enrichment is its ability to support the dynamic extension of resource meta-information by application and to provide means to share this meta-information with others. It implements the concept of groups of interest with access control on meta-information. The concept described in Matriona are in the process to be published.
- Little Thumb Base (LithBase) is an independent knowledge base that provides the same enrichment capabilities than Matriona but imposes fewer constraints on the architecture of applications. It is a shared database implemented on simple low power nodes (esp32) that are cheap to deploy, flash and use. The idea behind LithBase is to decouple the storage from the framework and to provide a standard mechanism to share information. Ultimately we want to use its capabilities to implement a registry in the manner of Consul with meta-information enrichment and sharing mechanisms. By focussing only on the discovery mechanism and information sharing, LithBase imposes fewer constraints on applications and comply more with the goal of being ready to use in existing applications. This is still a work in progress. This solution also raises the issue of trust and control over access to this information. It is indeed necessary for applications to be able to determine the source of the additional information and to determine who will have access to the information they add. We have also been experimenting with access control mechanism that is implemented by LithBase. We are currently using elliptic cryptography to allow private information sharing between groups. Ultimately the goal of this project is to produce a coordinating object that implements generic mechanisms favouring opportunistic behaviours of surrounding applications.

HYBRID Project-Team

7. New Results

7.1. Virtual Reality Tools and Usages

7.1.1. Gesture recognition for VR

Spatial and Rotation Invariant 3D Gesture Recognition Based on Sparse Representation

Participants: Ferran Argelaguet and Anatole Lécuyer

Advances in motion tracking technology, especially for commodity hardware, still require robust 3D gesture recognition in order to fully exploit the benefits of natural user interfaces. In this work [10], we introduced a novel 3D gesture recognition algorithm based on the sparse representation of 3D human motion. The sparse representation of human motion provides a set of features that can be used to efficiently classify gestures in real-time. Compared to existing gesture recognition systems, the proposed approach enables full spatial and rotation invariance and provides high tolerance to noise. Moreover, the proposed classification scheme takes into account the inter-user variability which increases gesture classification accuracy in user-independent scenarios. We validated our approach with existing motion databases for gestural interaction and performed a user evaluation with naive subjects to show its robustness to arbitrarily defined gestures. The results showed that our classification scheme has high classification accuracy for user-independent scenarios even with users who have different handedness. We believe that sparse representation of human motion will pave the way for a new generation of 3D gesture recognition systems in order to fully open the potential of natural user interfaces.

Figure 2. A participant interacting with the proposed gesture recognition system.

This work was done in collaboration with PANAMA team.

7.1.2. Automatic tools for the evaluation of VR systems

AGENT: Automatic Generation of Experimental Protocol Runtime

Participants: Gwendal Le Moulec, Ferran Argelaguet, Valérie Gouranton and Bruno Arnaldi

Due to the nature of Virtual Reality (VR) research, conducting experiments in order to validate the researchers' hypotheses is a must. However, the development of such experiments is a tedious and time-consuming task. We proposed in [18] to make this task easier, more intuitive and faster with a method able to describe and generate the most tedious components of VR experiments. The main objective is to let experiment designers focus on their core tasks: designing, conducting, and reporting experiments. To that end, we propose the use of Domain-Specific Languages (DSLs) to ease the description and generation of VR experiments. An analysis of published VR experiments is used to identify the main properties that characterize VR experiments. This allowed us to design AGENT (Automatic Generation of ExperimeNtal proTocols), a DSL for specifying and generating experimental protocol runtimes. AGENT allows experiment designers to design an Experimental Conditions Model (see Figure 3 -left) and a Protocol Model (see Figure 3 -right) in the AGENT editor. The models are then automatically compiled into code that can be integrated into VR development tools, e.g. Unity. We demonstrated the feasibility of our approach by using AGENT within two experiments.

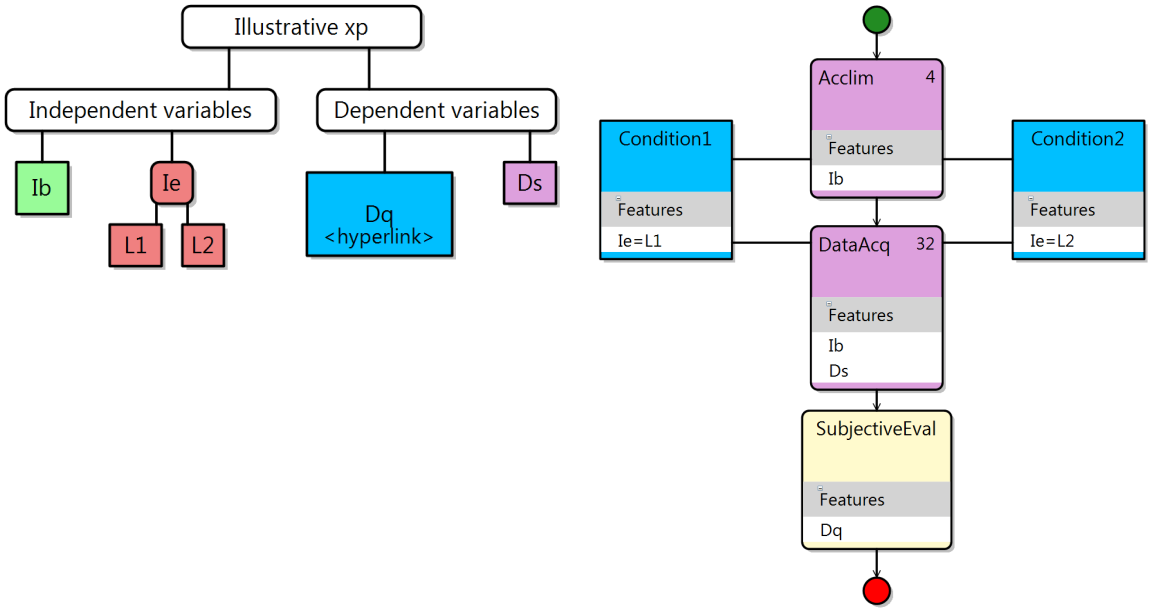

Figure 3. Examples of "Experimental Conditions" Model (left) and "Protocol" Model (right) which are both editable with the AGENT editor.

This work was done in collaboration with DIVERSE team.

7.1.3. Customer behavior and analyses in VR

The use of immersive Virtual Reality to investigate consumer perceptions and purchase behavior toward non-standard fruits and vegetables

Participants: Jean-Marie Normand and Guillaume Moreau

With the growth of organic Fruits and Vegetables (FaVs) markets, there is now a trend in marketing research toward studies of non-standardized fruits and vegetables in stores. Yet, because of the decaying nature of FaVs, it is difficult to conduct such studies. A solution is to conduct them within a Virtual Environment (VE) (with virtual FaVs). Therefore, it is of interest to develop an approach to generate a large variety and variability of FaVs, so one can later use them in a VE. First, we introduced a pipeline to generate a large variability of FaVs, focusing both on their shape and on their external appearance [29]. Regarding the shape, we use a semi-automated approach. A parametric Skeletal Structure and Generalized Cylinders (GCs) generates their overall shape and metaball-based techniques give them an organic aspect. Regarding their external appearance, we use a particle system approach to simulate their modifications over time. This particle system-based approach decomposes FaVs appearance changes into distinct visual characteristics producing different texture maps that can be combined.

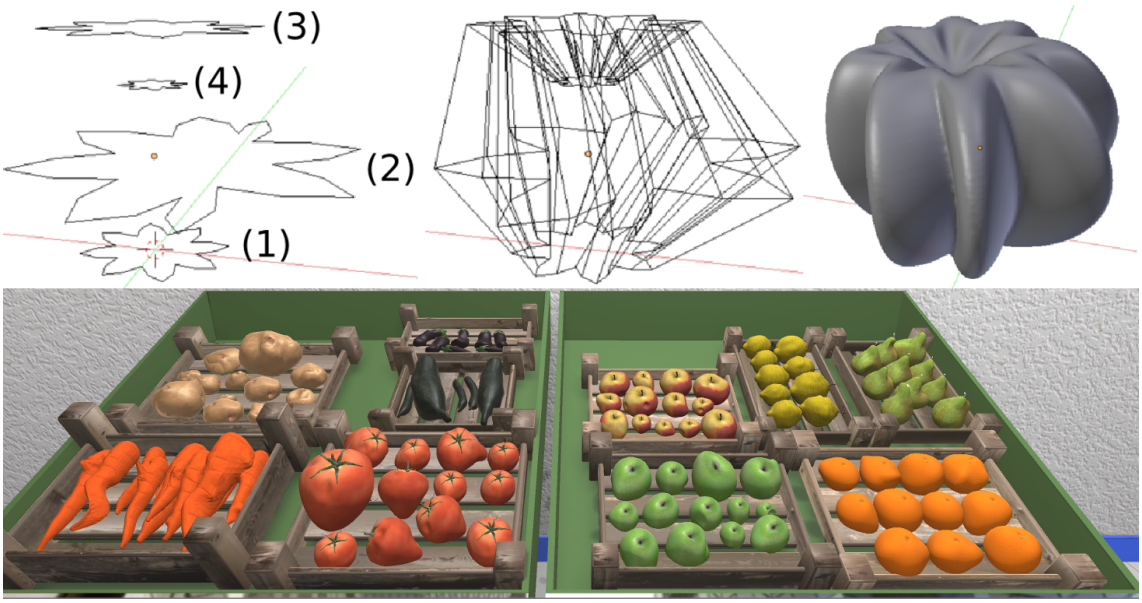

Figure 4. Our semi-automated process using a skeletal structure and cross-sections to generate different Fruits and Vegetables (Top). An example of semi-automatically generated FaVs with different levels of deformity (Bottom).

Second, we conducted an immersive virtual reality user study aimed at investigating how customers perceive and if they would purchase non standard (i.e. misshaped) fruits and vegetables (FaVs) in supermarkets and hypermarkets [23]. Indeed, food waste is a major issue for the retail sector and a recent trend is to reduce it by selling non-standard goods. An important question for retailers relates to the FaVs' " level of abnormality " that consumers would agree to buy. However, this question cannot be tackled using " classical " marketing techniques that perform user studies within real shops since fresh produce such as FaVs tend to rot rapidly preventing studies to be repeatable or to be run for a long time. In order to overcome those limitations, we created a virtual grocery store with a fresh FaVs section where 142 participants were immersed using an Oculus Rift DK2 HMD. Participants were presented either "normal", "slightly misshaped", "misshaped" or "severely misshaped" FaVs. Results show that participants tend to purchase a similar number of FaVs whatever their deformity. Nevertheless participants' perceptions of the quality of the FaV depend on the level of abnormality.

This work was done in collaboration with Audencia Business School, Nantes, France.

Figure 5. Our virtual supermarket, a participant and a close-up view of the participant on the Fruits and Vegetables booth.

7.2. Physically-Based Simulation and Haptic Feedback

7.2.1. Physically-based simulation

Elasticity-based Clustering for Haptic Interaction with Multi-Resolution Heterogeneous Deformable Objects

Participants: Benoît Le Gouis, Maud Marchal, Bruno Arnaldi and Anatole Lécuyer

Physically-based simulation of heterogeneous objects remains a strong computational challenge for many VR applications, especially when involving haptic interaction. In [17], we introduced a novel physically-based multi-resolution approach for haptic interaction with heterogeneous deformable objects. Our method called "Elasticity-based Clustering" is based on the clustering and aggregation of elasticity inside an object, so to create large homogeneous volumes preserving important features of the initial repartition. Such a creation of large and homogeneous volumes simplifies the attribution of elasticity to the elements of the coarser geometry. We could successfully implement and test our approach within a complete and real-time haptic interaction pipeline compatible with consumer-grade haptic devices. We evaluated the performance of our approach on a large set of elasticity configurations using a perception-based quality criterion. Our results show that for 90% of studied cases our method can achieve a 6 times speedup in the simulation time with no theoretical perceptual difference.

Real-time Target Tracking of Soft Tissues in 3D Ultrasound Images Based on Robust Visual Information and Mechanical Simulation

Participants: Maud Marchal

In [], we presented a real-time approach that allows tracking deformable structures in 3D ultrasound sequences. Our method consists in obtaining the target displacements by combining robust dense motion estimation and mechanical model simulation. We performed evaluation of our method through simulated data, phantom data, and real-data. Results demonstrated that this novel approach has the advantage of providing correct motion estimation regarding different ultrasound shortcomings including speckle noise, large shadows and ultrasound

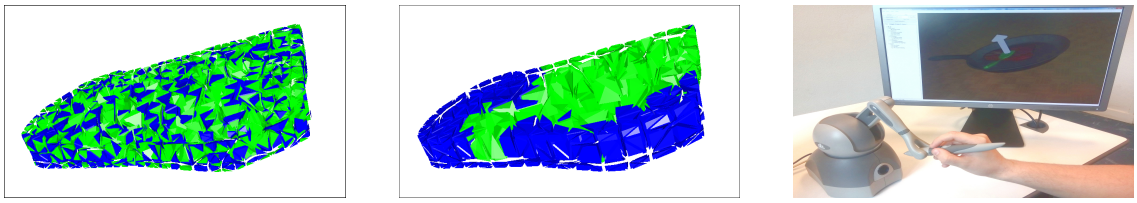

Figure 6. Overview of our elasticity-based clustering approach: (left) Based on an heterogeneous object composed of elements with different elasticity values, (center) we propose to build elasticity clusters to improve the computation time performances, (right) thus allowing haptic interaction while keeping similar perceptual sensations.

gain variation. Furthermore, we could show the good performance of our method with respect to state-of-the-art techniques by testing on the 3D databases provided by MICCAI CLUST'14 and CLUST'15 challenges.

This work was done in collaboration with Lagadic team and IRT B-Com.

7.2.2. Haptic feedback

Haptic Rendering of FEM-based Tearing Simulation using Clusterized Collision Detection

Participants: Benoît Le Gouis, François Lehericey, Maud Marchal, Valérie Gouranton, Bruno Arnaldi and Anatole Lécuyer

Haptic rendering of deformable phenomena remains computationally-demanding, especially when topology modifications are simulated. Within this context, the haptic rendering of tearing phenomena is under-explored as of today. In [16] we proposed a fully-functional interaction pipeline for physically-based simulation of deformable surface tearing allowing to reach haptic interactive rates. It relies on a high efficiency collision detection algorithm for deformable surface meshes, combined with an efficient FEM-based simulation of deformable surfaces enabling tearing process. We especially introduced a novel formulation based on clusters for the collision detection to improve computation time performances. Our approach was illustrated through interactive use-cases of tearing phenomena with haptic feedback, showing its ability to handle realistic rendering of deformable surface tearing on consumer-grade haptic devices.

FlexiFingers: Multi-Finger Interaction in VR Combining Passive Haptics and Pseudo-Haptics

Participants: Maud Marchal, Benoît Le Gouis, Ferran Argelaguet and Anatole Lécuyer

3D interaction in virtual reality often requires to manipulate and feel virtual objects with our fingers. Although existing haptic interfaces can be used for this purpose (e.g. force-feedback exoskeleton gloves), they are still bulky and expensive. We introduced a novel multi-finger device called "FlexiFingers" that constrains each digit individually and produces elastic forcefeedback [9]. FlexiFingers leverages passive haptics in order to offer a lightweight, modular, and affordable alternative to active devices. Moreover, we combined Flexifingers with a pseudo-haptic approach that simulates different levels of stiffness when interacting with virtual objects. We illustrated how this combination of passive haptics and pseudo-haptics could benefit multi-finger interaction through several use cases related to music learning and medical training. These examples suggest that our approach could find applications in various domains that require an accessible and portable way of providing haptic feedback to the fingers.

7.3. Augmented Reality

7.3.1. Perception in augmented reality

AR Feels "Softer" than VR: Haptic Perception of Stiffness in Augmented versus Virtual Reality

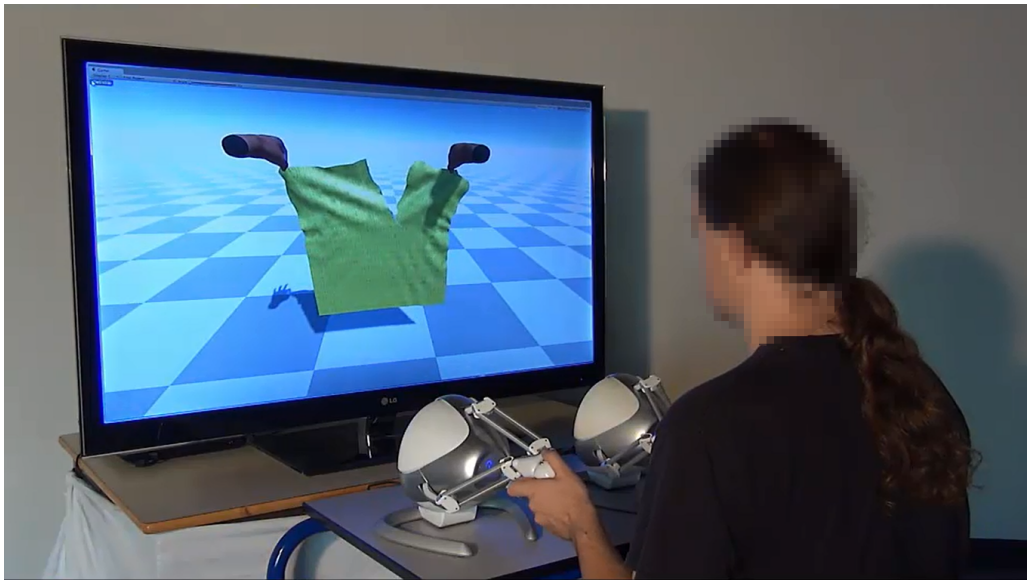

Figure 7. Our FEM-based method allows for the bimanual haptic tearing of deformable surfaces.

Figure 8. The FlexiFingers is a multi-finger device, combined with a pseudo-haptic approach, that can be used in music learning applications for instance.

Participants: Yoren Gaffary, Benoît Le Gouis, Maud Marchal, Ferran Argelaguet, Anatole Lécuyer and Bruno Arnaldi

Does it feel the same when you touch an object in Augmented Reality (AR) or in Virtual Reality (VR)? In [3] we studied and compared the haptic perception of stiffness of a virtual object in two situations: (1) a purely virtual environment versus (2) a real and augmented environment. We have designed an experimental setup based on a Microsoft HoloLens and a haptic force-feedback device, enabling to press a virtual piston, and compare its stiffness successively in either Augmented Reality (the virtual piston is surrounded by several real objects all located inside a cardboard box) or in Virtual Reality (the same virtual piston is displayed in a fully virtual scene composed of the same other objects). We have conducted a psychophysical experiment with 12 participants. Our results show a surprising bias in perception between the two conditions. The virtual piston is on average perceived stiffer in the VR condition compared to the AR condition. For instance, when the piston had the same stiffness in AR and VR, participants would select the VR piston as the stiffer one in 60% of cases. This suggests a psychological effect as if objects in AR would feel ”softer” than in pure VR. Taken together, our results open new perspectives on perception in AR versus VR, and pave the way to future studies aiming at characterizing potential perceptual biases.

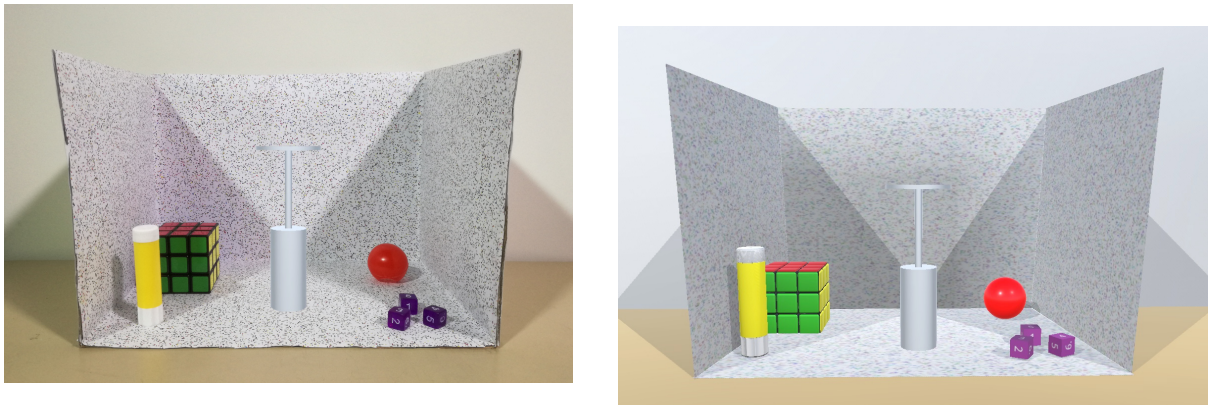

Figure 9. In our experiment, participants could interact with a virtual piston superimposed inside a real cardboard box in AR (left), and with the same piston inside a virtual replica of the box in VR (right).

7.3.2. Interaction in augmented reality

Evaluation of Facial Expressions as an Interaction Mechanism and their Impact on Affect, Workload and Usability in an AR game

Participants: Jean-Marie Normand and Guillaume Moreau

With the recent development of Head Mounted Display (HMD) for Virtual Reality (VR) allowing to track and recognize user's Facial Expression (FE)s in real-time, we investigated the impact that the use of FEs as an action-trigger input mechanism (e.g. a FE mapped to a single action) has on our emotional state; as well as their workload and usability compared to the use of a controller button. In [22] we developed an Augmented Reality (AR)-based memory card where the users select virtual cards using a wand and flip them using either a FE (smiling; frowning) or a Xbox controller button. The users were separated into three groups: (1) flipping the card with a smile (n = 10); (2) flipping the card with a frown (n = 8) and (3) flipping the cards with the Xbox controller button (n = 11). We did not see any significant differences between our groups in: (i) the positive and negative affect of the participants and (ii) the reported workload and usability, thus highlighting that the FEs could be used inside a HMD in the same way as a controller button.

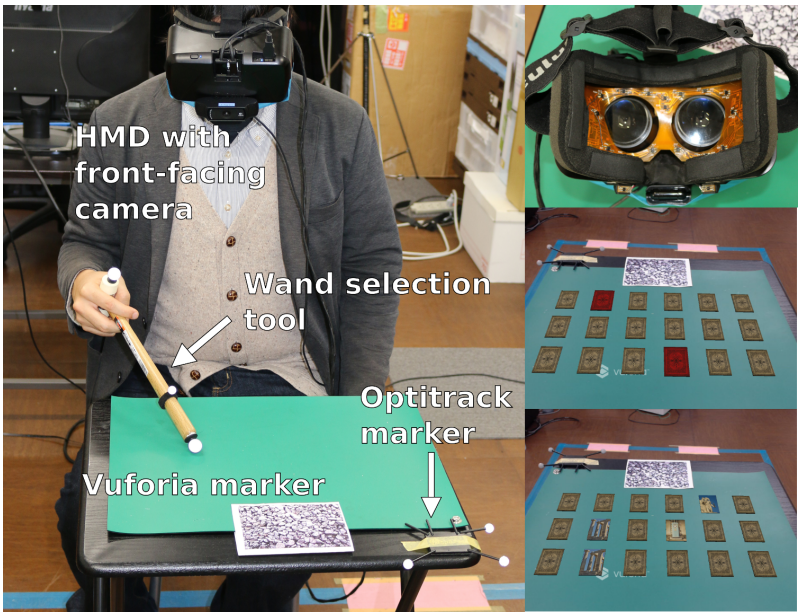

Figure 10. A participant wearing the Expression-Wear and playing our AR memory card game. Top right: The ExpressionWear embedded in the Oculus Rift. Middle and Bottom right: two states of our AR memory card game.

This work was done in collaboration with the Interactive Media Lab of Keio University (Japan).

A Sate-of-the-Art on the Combination of Brain-Computer Interfaces and Augmented Reality

Participants: Hakim Si-Mohammed, Ferran Argelaguet and Anatole Lécuyer

We have reviewed the state-of-the art of using Brain-Computer Interfaces in combination with Augmented Reality (AR) [21]. In this work, first we introduced the field of AR and its main concepts. Second, we described the various systems designed so far combining AR and BCI categorized by their application field: medicine, robotics, home automation and brain activity visualization. Finally, we summarized and discussed the results of our survey, showing that most of the previous works made use of P300 or SSVEP paradigms with EEG in Video See-Through systems, and that robotics is a main field of application with the highest number of existing systems.

This work was done in collaboration with MJOLNIR team.

7.3.3. Tracking

Increasing Optical Tracking Workspace of VR Applications using Controlled Cameras

Participants: Guillaume Cortes, Anatole Lécuyer

We have proposed an approach to greatly increase the tracking workspace of VR applications without adding new sensors [14]. Our approach relies on controlled cameras able to follow the tracked markers all around the VR workspace providing 6DoF tracking data. We designed the proof-of-concept of such approach based on two consumer-grade cameras and a pan-tilt head. The resulting tracking workspace could be greatly increased depending on the actuators' range of motion. The accuracy error and jitter were found to be rather limited during camera motion (resp. 0.3cm and 0.02cm). Therefore, whenever the final VR application does not require a perfect tracking accuracy over the entire workspace, we recommend using our approach in order to enlarge the tracking workspace.

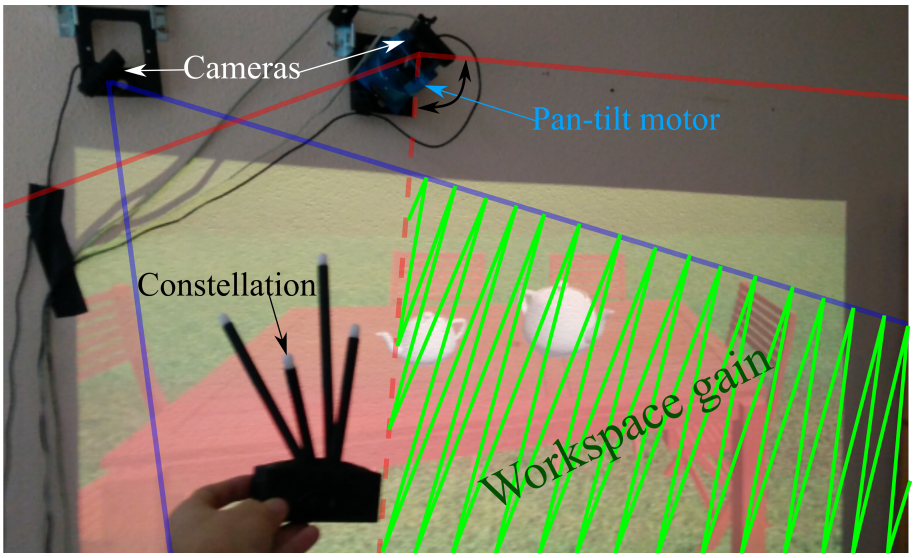

Figure 11. Our optical tracking prototype based on controlled cameras illustrated on a wall-sized VR application.

This work was done in collaboration with LAGADIC team.

An Optical Tracking System based on Hybrid Stereo/Single-View Registration and Controlled Cameras

Participants: Guillaume Cortes, Anatole Lécuyer

Optical tracking is also widely used in robotics applications such as unmanned aerial vehicle (UAV) localization. Unfortunately, such systems require many cameras and are, consequently, expensive. We proposed an approach to increase again the optical tracking volume without adding cameras [13]. First, when the target becomes no longer visible by at least two cameras we propose a single-view tracking mode which requires only one camera. Furthermore, we propose to rely again on controlled cameras able to track the UAV all around the volume to provide 6DoF tracking data through multi-view registration. This is achieved by using a visual servoing scheme. The two methods can be combined in order to maximize the tracking volume. We propose a proof-of-concept of such an optical tracking system based on two consumer-grade cameras and a pan-tilt actuator and we used this approach on UAV localization.

This work was done in collaboration with LAGADIC team.

7.4. Brain-Computer Interfaces

7.4.1. BCI methods and techniques

Designing Guiding Systems for BCI

Participants: Nataliya Kosmyna and Anatole Lécuyer

The Brain–Computer Interface (BCI) community has focused the majority of its research efforts on signal processing and machine learning, mostly neglecting the human in the loop. Guiding users on how to use a BCI is crucial in order to teach them to produce stable brain patterns. In [5] we explored the instructions and feedback for BCIs in order to provide a systematic taxonomy to describe the BCI guiding systems. The purpose of our work was to give necessary clues to the researchers and designers in Human–Computer Interaction (HCI) in making the fusion between BCIs and HCI more fruitful but also to better understand the possibilities BCIs can provide to them.

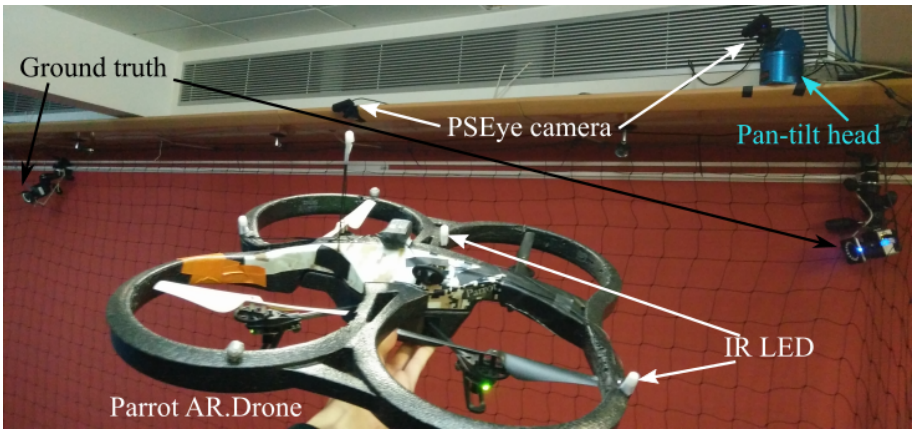

Figure 12. Our optical tracking prototype based on controlled cameras and hybrid stereo/single-view registration. The tracking is used for UAV indoor localization purposes.

Towards Understanding Inverse Models in BCI

Participants: Jussi Lindgren

In the scope of the LABEX CominLabs project "SABRE", we have investigated the applicability of physiology-based source reconstruction for Brain-Computer Interfaces. The BCI interfaces leave a lot to be desired in terms of their accuracy and speed. Can source reconstruction help? We explained how the source reconstruction techniques relate to the currently mainstream machine learning methods that may recover the sources implicitly [6]. We explained the different approaches in a common linear dictionary framework and review the different ways to obtain the dictionary parameters. Our analysis suggests physiological source reconstruction may improve BCI accuracy if machine learning is not used or where it produces less optimal parameters. We considered the effect of source reconstruction on some major difficulties in BCI classification, namely information loss, feature selection and nonstationarity of the EEG. The provided analysis and discussion should help in understanding, applying, comparing and improving such techniques in the future.

Cognitive Demand of BCI

Participants: Andeol Evain, Ferran Argelaguet and Anatole Lécuyer

BCIs are presumably supposed to require the full attention of their users and to lose accuracy if they pay attention to another task. This assertion has been verified with several BCI paradigms (e.g. P300). But the cognitive demand of the promising SSVEP paradigm had never been specifically assessed yet. In [15] we measured the accuracy of an SSVEP-based BCI used by 26 participants in various conditions of mental workload. Our analysis revealed that surprisingly, for this type of BCI, little attention is actually needed from participants to reach optimal accuracy: participants were able to successfully perform a complex secondary task (N-back) without degrading the BCI accuracy. The same observation was made whether visual or auditive attention was solicited. These results indicate that SSVEP is a low-demanding paradigm in terms of cognitive resources, and are encouraging for its use in complex interaction settings.

This work was done in collaboration with MJOLNIR team.

7.4.2. Neurofeedback

How to Build a Hybrid Neurofeedback Platform Combining EEG and fMRI

Participants: Marsel Mano, Lorraine Perronnet and Anatole Lécuyer

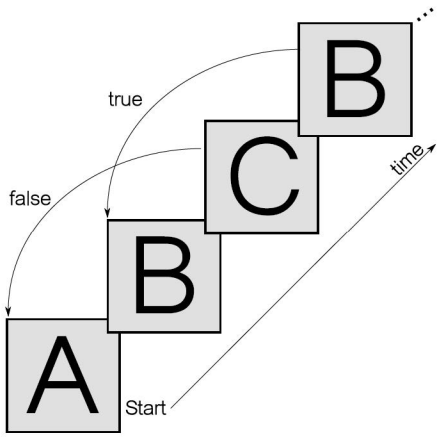

Figure 13. N-back task used in our study to control the level of difficuly and cognitive workload.

Multimodal neurofeedback estimates brain activity using information acquired with more than one neurosignal measurement technology. We have studied and described how to set up and use a hybrid platform based on simultaneous electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), then we illustrated how to use it for conducting bimodal neurofeedback experiments in [20]. This work is intended for those willing to build a multimodal neurofeedback system, to guide them through the different steps of the design, setup, and experimental applications, and help them choose a suitable hardware and software configuration. Furthermore, we reported practical information from bimodal neurofeedback experiments conducted in our lab (see Figure 14). The platform that we presented has a modular parallel processing architecture that promotes real-time signal processing performance and simple future addition and/or replacement of processing modules. Various unimodal and bimodal neurofeedback experiments conducted in our lab showed high performance and accuracy. Currently, the platform is able to provide neurofeedback based on electroencephalography and functional magnetic resonance imaging, but the architecture and the working principles described here are valid for any other combination of two or more real-time brain activity measurement technologies.

This work was done in collaboration with VISAGES team.

Unimodal Versus Bimodal EEG-fMRI Neurofeedback of a Motor Imagery Task

Participants: Lorraine Perronnet and Anatole Lécuyer

Neurofeedback is a promising tool for brain rehabilitation and peak performance training. Neurofeedback approaches usually rely on a single brain imaging modality such as EEG or fMRI. Combining these modalities for neurofeedback training could allow to provide richer information to the subject and could thus enable him/her to achieve faster and more specific self-regulation. Yet unimodal and multimodal neurofeedback have never been compared before. In [8] we introduced a simultaneous EEG-fMRI experimental protocol in which participants performed a motor-imagery task in unimodal and bimodal NF conditions (see Figure 15). With this protocol we were able to compare for the first time the effects of unimodal EEG-neurofeedback and fMRI-neurofeedback versus bimodal EEG-fMRI-neurofeedback by looking both at EEG and fMRI activations. We also proposed a new feedback metaphor for bimodal EEG-fMRI-neurofeedback that integrates both EEG and fMRI signal in a single bi-dimensional feedback (a ball moving in 2D). Such a feedback is intended to relieve the cognitive load of the subject by presenting the bimodal neurofeedback task as a single regulation task instead of two. Additionally, this integrated feedback metaphor gives flexibility on defining

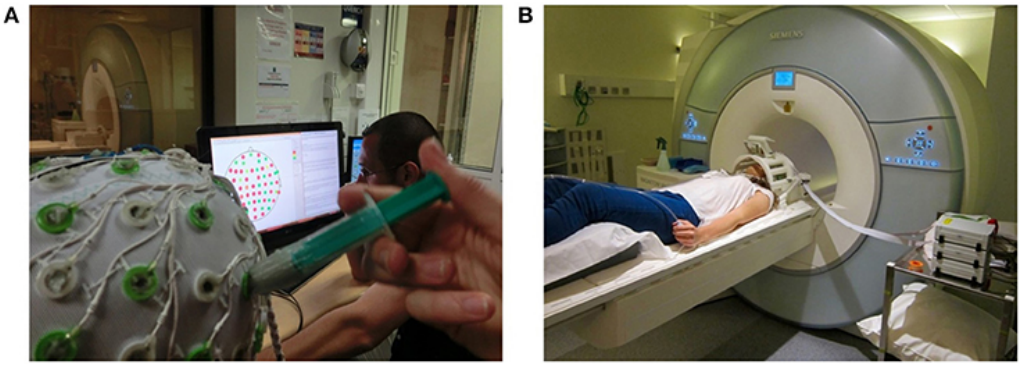

Figure 14. Experimental platform: (A) EEG subsystem installation outside the MR room, (B) installation of the MR coil.

a bimodal neurofeedback target. Participants were able to regulate activity in their motor regions in all NF conditions. Moreover, motor activations as revealed by offline fMRI analysis were stronger during EEG-fMRI-neurofeedback than during EEG-neurofeedback. This result suggests that EEG-fMRI-neurofeedback could be more specific or more engaging than EEG-neurofeedback. Our results also suggest that during EEG-fMRI-neurofeedback, participants tended to regulate more the modality that was harder to control. Taken together our results shed first light on the specific mechanisms of bimodal EEG-fMRI-neurofeedback and on its added-value as compared to unimodal EEG-neurofeedback and fMRI-neurofeedback.

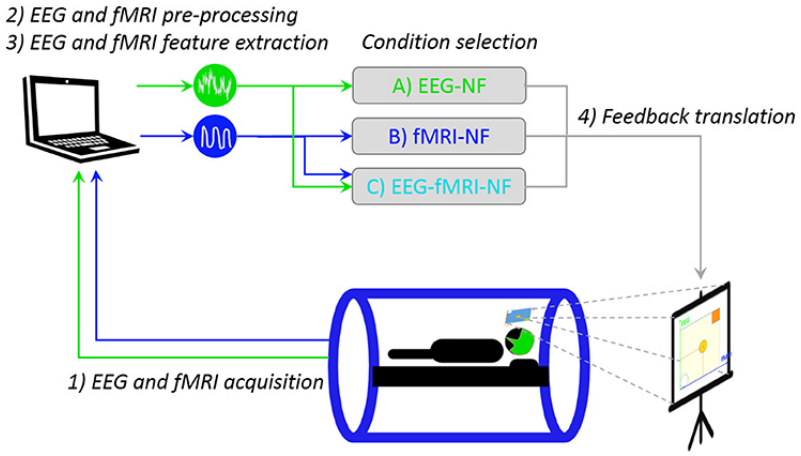

Figure 15. Real-time multimodal EEG/fMRI Neurofeedback experiment. The participant is lying in the MR tube with a 64-channel MR-compatible EEG cap. EEG and fMRI are simultaneously acquired then pre-processed with custom Matlab code. The EEG and fMRI laterality features are computed and eventually translated as a displacement of the ball on the stimulation screen, the image of which is projected on the mirror mounted on the head coil.

This work was done in collaboration with VISAGES team.

Investigating Neurophysiological Correlates of Covert Attention in Soccer Goalkeepers

Participants: Camille Jeunet, Ferran Argelaguet and Anatole Lécuyer

Soccer goalkeepers must process information from their peripheral vision at the same time they look towards the ball. This ability, committing attention to a position other than the fixation point, is called Covert Visuo-Spatial Attention or CVSA. CVSA being essential to reach high performances, it is primordial to find innovative and efficient ways of improving it. Neurofeedback, which consists in training specific brain features in order to enhance a cognitive ability, has been proven to increase attentional abilities. Also, different studies have suggested the existence of a neurophysiological marker specific to covert attention: a lateralised modulation of the alpha waves in the visual cortex. Moreover, it has been shown possible to compute this marker online, thus opening the door to a potential neurofeedback training procedure. In this view, we have proposed in a first instance to further investigate the relevance of this marker for soccer goalkeepers. The objective was here to answer the following questions: Is this marker transferrable to goalkeepers? How stable is it across athletes? Does it depend on their expertise?
This work was presented at the World Conference on Science and Soccer (Rennes, France, May 2017). It was done in collaboration with M2S Laboratory and EPFL.

7.5. Cultural Heritage

7.5.1. VR and AR tools for cultural heritage

EvoluSon: Walking Through an Interactive History of Music

Participants: Ronan Gaugne, Florian Nouviale and Valérie Gouranton

The EvoluSon project [4] proposes an immersive experience where the spectator explores an interactive visual and musical representation of the main periods of the history of Western music. The musical content is constituted of original musical compositions based on the theme of Bach's Art of Fugue to illustrate eight main musical eras from Antiquity to the contemporary epoch. The EvoluSon project contributes at the same time to the usage of VR for intangible culture representation and to interactive digital art that puts the user at the centre of the experience. The EvoluSon project focuses on music through a presentation of the history of Western music, and uses virtual reality to valorise the different pieces through the ages. The user is immersed in a coherent visual and sound environment and can interact with both modalities (see Figure 16).

This project was done in collaboration with the Research Laboratory on Art and Music of University Rennes 2.

Immersive Point Cloud Manipulation for Cultural Heritage Documentation

Participants: Jean-Baptiste Barreau, Ronan Gaugne and Valérie Gouranton

Virtual reality combined with 3D digitisation allows to immerse archaeologists in 1:1 copies of monuments and sites. However, scientific communication of archaeologists is based on 2D representations of the monuments they study. In [2] we proposed a virtual reality environment with an innovative cutting-plan tool to dynamically produce 2D cuts of digitized monuments. A point cloud is the basic raw data obtained when digitizing cultural heritage sites or monuments with laser scans or photogrammetry. These data represent a rich and faithful record provided that they have adequate tools to exploit them. Their current analyses and visualizations on PC require software skills and can create ambiguities regarding the architectural dimensions. We conceived a toolbox to explore and manipulate such data in an immersive environment, and to dynamically generate 2D cutting planes usable for cultural heritage documentation and reporting (see Figure 17).

MAAP Annotate: When Archaeology meets Augmented Reality for Annotation of Megalithic Art

Participants: Jean-Marie Normand

Figure 16. EvoluSon application in Immersia, Middle-Age era.

Figure 17. Immersive manipulation of the point cloud of the "Salle du Jeu de Paume" of Rennes, in Immersia room.

Megalithic art is a spectacular form of symbolic representation found on prehistoric monuments. Carved by Europe's first farmers, this art allows an insight into the creativity and vision of prehistoric communities. As examples of this art continue to fade, it is increasingly important to document and study these symbols. In [11] we introduced MAAP Annotate, a Mixed Reality annotation tool from the Megalithic Art Analysis Project (MAAP). It provides an innovative method of interacting with megalithic art, combining cross-disciplinary research in digital heritage, 3D scanning and imaging, and augmented reality. The development of the tool is described, alongside the results of an evaluation carried out on a group of archaeologists from University College Dublin, Ireland. It is hoped that such tools will enable archaeologists to collaborate worldwide, and non-specialists to experience and learn about megalithic art (see Figure 18).

This work was done in collaboration with the School of Computer Science and Informatics and the School of Archaeology of University College Dublin (UCD).

Figure 18. A real Irish megalith engraved with petroglyphs (Top-left) and a participant using the HoloLens to annotate the 3D scan of the megalith (Top-right). Overview of the MAAP Annotate user interface (Middle row), and two examples of manual AR annotations (Bottom row).

7.5.2. Multi-modal images and 3D printing for cultural heritage

Physical Digital Access Inside Archaeological Material

Participants: Théophane Nicolas, Ronan Gaugne and Valérie Gouranton

Traditionally, accessing the interior of an artefact or an archaeological material is a destructive activity. We proposed an alternative non-destructive technique, based on a combination of medical imaging and advanced transparent 3D printing. Our approach proposes combining a computed tomography (CT) scan and advanced

3D printing to generate a physical representation of an archaeological artefact or material [7]. This project is conducted with archaeologists from Inrap and computer scientists from Inria-IRISA. The goal of the project is to propose innovative practices, methods and tools for archaeology based on 3D digital techniques. We notably proposed a workflow (see Figure 19) where the CT scan images are used to produce volume and surface 3D data which serve as a basis for new evidence usable by archaeologists. This new evidence can be observed in interactive 3D digital environments or through physical copies of internal elements of the original material.

This work was done in collaboration with Inrap.

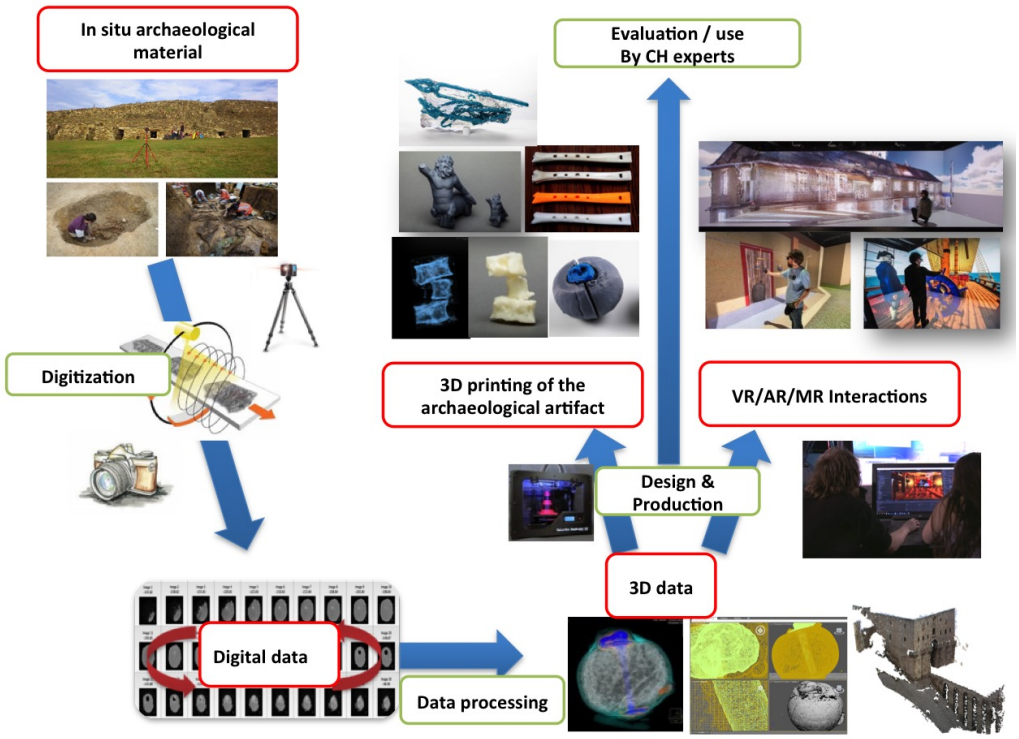

Figure 19. Our workflow for combining a computed tomography (CT) scan and advanced 3D printing to generate a physical representation of an archaeological artefact.

Combining CT-scan, Photogrammetry, 3D Printing and Mixed Reality

Participants: Théophane Nicolas, Ronan Gaugne, Bruno Arnaldi and Valérie Gouranton

Archaeological artefacts and the sediments that contain them constitute the sometimes tenuous evidence that requires analysis, preservation and showcasing. Different methods of digital analysis that provide non destructive solutions to preserve, analyse and showcase archaeological heritage have been developed over recent years. However these techniques are often restricted to the visible surface of the objects, monuments or sites. The techniques used in medical imaging are more and more frequently used in archaeology as they give non destructive access to the artefacts' internal and often fragile structure. This use is mostly limited to a simple visualisation. The information obtained by CT-scan is transcribed in a visual manner and its inherent detail can be used much more widely in the domain of the latest 3D technologies such as virtual reality, augmented reality, multimodal interactions and additive manufacturing. In combining these medical imaging techniques, it becomes possible to identify and scientifically analyse by efficient and non destructive methods non visible

objects, to assess their fragility and their state of preservation. It is also possible to assess the restoration of a corroded artefact, to visualise, to analyse and to physically manipulate an inaccessible or fragile object (CT, 3D printing) and to observe the context of our hidden archaeological heritage (virtual reality, augmented reality or mixed, 3D). The development of digital technologies will hopefully lead to a democratisation of this type of analysis. We could illustrate our approach using the study of several artefacts from the recent excavation of the Warcq chariot burial (Ardennes, France). We notably presented in [28] a physical interaction with inaccessible objects based on a transparent 3D printing of a horse's cranium (see Figure 20).

This work was done in collaboration with Inrap, Image ET and University Paris 1.

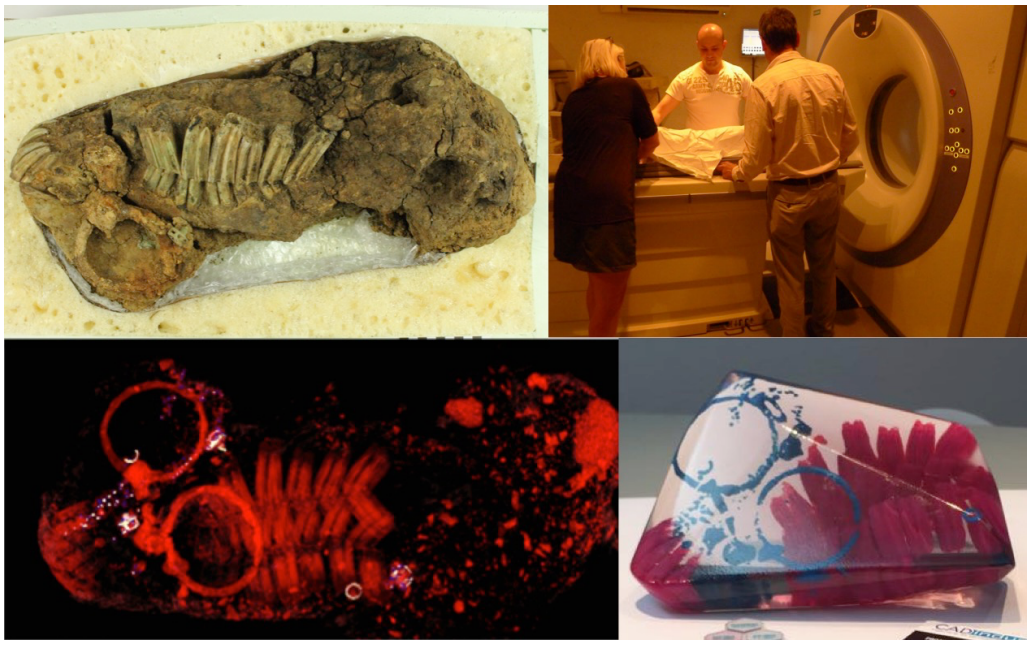

Figure 20. Original horse cranium (Top-left) and CT-scan of the cranium (Top-right), volume rendering from the CT scan (Bottom-left) and then transparent 3D printing of the same object (Bottom-right).

A Multimodal Digital Analysis of a Mesolithic Clavicle: Preserving and Studying the Oldest Human Bone in Brittany

Participants: Jean-Baptiste Barreau, Ronan Gaugne and Valérie Gouranton

The oldest human bone of Brittany was dug up from the mesolithic shell midden of Beg-er-Vil in Quiberon and dated about 8200 years. The low acid soils of these dump area represent exceptional sedimentary conditions. For these reasons, but also because these bones have a very particular patrimonial and symbolic value, their study goes altogether with concerns of conservation and museographic presentation. The clavicle is constituted of two pieces discovered separately at a one meter distance from each other. The two pieces match, so it can be assemble in a single fragment of approximately 7 centimeters. Cut-marks are clearly visible on the surface of these bones. They are bound to post- mortem processing which it is necessary to better qualify. The clavicle was studied through a process that combines advanced 3D image acquisition, 3D processing, and 3D printing with the goal to provide relevant support for the experts involved [24]. The bones were first scanned with a CT scan, and digitized with photogrammetry in order to get a high quality textured model. The CT scan appeared to be insufficient for a detailed analysis. The study was thus completed with a μCT providing a very accurate 3D model of the bone. Several 3D-printed copies of the collarbone were produced in order to

constitute tangible support easy to annotate for sharing between the experts involved in the study. The 3D models generated from μCT and photogrammetry, were combined to provide an accurate and detailed 3D model. This model was used to study desquamation and the different cut marks. These cut marks were also studied with traditional binoculars and digital microscopy. This last technique allowed characterizing the cut marks, revealing a probable meat cutting process with a flint tool (see Figure 21). This work of crossed analyses allowed to document a fundamental patrimonial piece, and to ensure its preservation.

This work was done in collaboration with Inrap, UMR CReAAH, CNRS-INE and Université Paris 1.

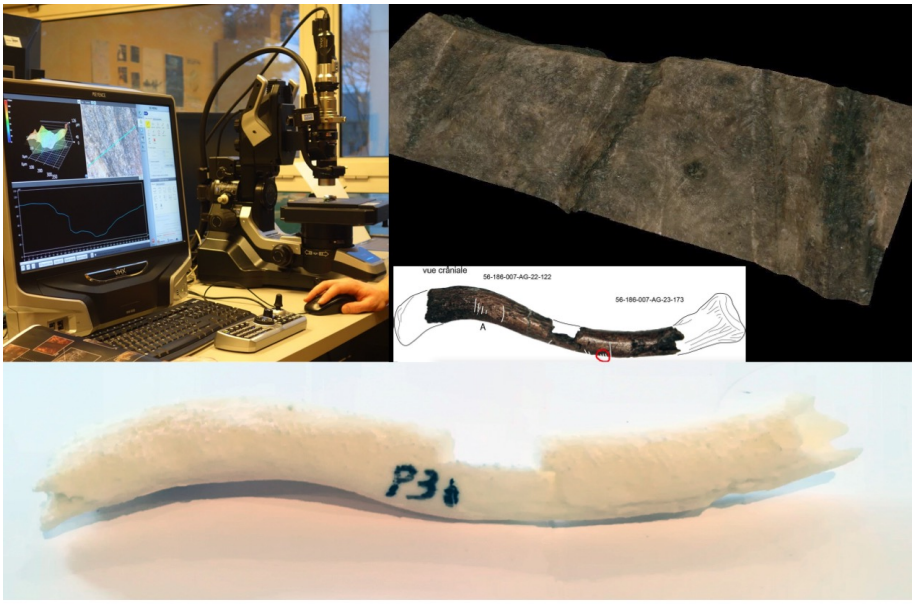

Figure 21. Digital microscopy of the clavicle (Top-left), and then detail of the cut marks in the digital model (Top-right), and annotated 3D printed clavicle (Bottom).

7.5.3. Generating 3D data for cultural heritage

3D Reconstruction of the Fortified Entrance of the Citadel of Aleppo from a few Sightseeing Photos

Participants: Jean-Baptiste Barreau and Ronan Gaugne

Built at the beginning of the 16th century by the final Mamluk sultan Al-Achraf Qânsûh Al-Ghûrî, the entrance to the Citadel of Aleppo was particularly affected by an earthquake in 1822, bombings during the Battle of Aleppo in August 2012, and a collapse of ramparts due to an explosion in July 2015. Even if compared to other Syrian sites, there are still enough vestiges to grasp the initial architecture, the civil war situation makes extremely difficult any "classic" process of digitization by photogrammetry or laser scanning. On this basis, we proposed in [25] a process to produce a 3D model "as relevant as possible" only from a few sightseeing photographs. This process combines fast 3D sketching by photogrammetry, 3D modeling and texture mapping and relies on a corpus based on pictures available on the net. Furthermore, it has the advantage to be applicable to destroyed monuments if sufficient pictures are available (see Figure 22).

This work was done in collaboration with UMR CReAAH and Inrap.

Figure 22. Rendering of the 3D model of Aleppo entrance.

Raising the Elevations of a Megalithic Architecture: Methodological Developments

Participants: Jean-Baptiste Barreau, Quentin Petit and Ronan Gaugne

Elevations have been little studied during explorations of megalithic architectures. For the past ten years, interest in these elevations has been growing in western France, particularly with the application of archeology of buildings and its tools to study them. Megalithic architecture, however, has its own characteristics that make manual surveys difficult. The first step presented in [26] was to acquire a 3D model, precise and manageable, of the whole architecture. Photogrammetry was tested, however the small space made it difficult to photograph. Laser scanner scanning has therefore been preferred. From the cloud of points obtained, a computer processing protocol was developed in order to obtain 2D images of the elevations according to the desired views. On these, a stone-to-stone design is possible in the laboratory and rectifiable directly in the field thanks to the use of a tablet computer. Our method has therefore met accessibility constraints. Above all, it allowed to increase the time devoted to the observation of ground, with a final result identical to a manual survey.

This work was done in collaboration with UMR CReAAH and SED Rennes.

Figure 23. Inside the cairn of Barnenez in Immersia room.

LACODAM Project-Team

7. New Results

7.1. Introduction

In this section, we organize our contributions this year along two of our research axes, namely Pattern Mining and Decision Support. These correspond to the contributions that has been accepted for publication this year.

7.1.1. Pattern Mining

In the domain of pattern mining we can categorize our contributions along the following lines:

- *Mining of novel types of patterns.* This includes temporal pattern mining, signature mining, opinion mining in uncertain databases, interval rules, and top-k item-centric mining. All these contributions have been proposed as solutions to problems in the domains of pharmaco-epidemiology, retail databases, biomedical databases, and analysis of speech corpora. We provide more details about these results in Sections 7.2 to 7.9 .
- *Data Mining with ASP.* Answer Set Programming is a powerful search tool in combinatorial spaces, which can be naturally ported to pattern mining, as the latter is a specific type of search problem. Our contributions include the application of ASP in the discovery of frequent, constrained, condensed, and rare sequential patterns. Sections 7.11 and 7.12 elaborate on our new research insights.
- *Data Mining for the masses.* In [14], we propose a communication model that bridges knowledge delivery between data miners and domain users in the field of library science. Our model proposes a five-steps process in order to achieve effective knowledge synthesis and delivery of insights to the domain users.

7.1.2. Decision Support

In regards to the axis of decision support, our contributions can be organized in two categories: exploration and diagnosis.

- *Exploration.* We propose two exploration methods in the context of analysis of trajectories and agro-environmental systems. We propose customized data models and resort to data-warehousing and multidimensional data representations to facilitate the querying, and thus the exploration and understanding of the data, for the sake of decision making. Our results in this line are further detailed in Sections 7.13 to 7.15 .
- *Diagnosis.* In Section 7.16 we propose a novel method for anomaly detection in time series by resorting to Extreme Value Theory. In addition, [21] offers a formalization of diagnosis based on automata with focus on discrete event systems.

7.2. Discriminant chronicles mining: Application to care pathways analytics

Participants: Yann Dauxais, Thomas Guyet, David Gross-Amblard [Druid], André Happe [Brest University Hospital/REPERES].

Pharmaco-epidemiology (PE) is the study of uses and effects of drugs in well defined populations. As medico-administrative databases cover a large part of the population, they have become very interesting to carry PE studies. Such databases provide longitudinal care pathways in real condition containing timestamped care events, especially drug deliveries. Temporal pattern mining becomes a strategic choice to gain valuable insights about drug uses. We propose DCM [8], [7], a new discriminant temporal pattern mining algorithm. It extracts chronicle patterns that occur more in a studied population than in a control population. We present satisfactory results on the identification of possible associations between hospitalizations for seizure and anti-epileptic drug switches in care pathway of epileptic patients.

A stable release of the DCM algorithm (see Section 6.5) have been deposed to the Program Protection Agency (APP) and is available online.

7.3. Purchase Signatures of Retail Customers

Participants: Clément Gautrais, Peggy Cellier [SemLis], Thomas Guyet, René Quiniou, Alexandre Termier.

In the retail context, there is an increasing need for understanding individual customer behavior in order to personalize marketing actions. We propose the novel concept of customer signature, that identifies a set of important products that the customer refills regularly [10]. Both the set of products and the refilling time periods give new insights on the customer behavior. Our approach is inspired by methods from the domain of sequence segmentation, thus benefiting from efficient exact and approximate algorithms. Experiments on a real massive retail dataset show the applicability of the signatures for understanding individual customers.

7.4. Topic Signatures in Political Campaign Speeches

Participants: Clément Gautrais, Peggy Cellier [SemLis], René Quiniou, Alexandre Termier.

Highlighting the recurrence of topics usage in candidates speeches is a key feature to identify the main ideas of each candidate during a political campaign. In this study [9], we develop a method combining standard topic modeling with signature mining for analyzing topic recurrence in speeches of Clinton and Trump during the 2016 American presidential campaign. The results show that the method extracts automatically the main ideas of each candidate and, in addition, provides information about the evolution of these topics during the campaign.

7.5. Expert Opinion Extraction from a Biomedical Database

Participants: Ahmed Samet, Thomas Guyet, Benjamin Négrevergne, Tien-Tuan Dao, Tuan Nha Hoang, Marie-Christine Ho Ba Tho.

This work tackles the problem of extracting frequent opinions from uncertain databases. This problem is encountered in real-world applications, such as the opinions of medical experts to evaluate the reliability level of biomedical data. We introduce the foundation of an opinion mining approach with the definition of pattern and support measure. The support measure is derived from the commitment definition. In [15], we proposed a new algorithm called OPMINER that extracts the set of frequent opinions modeled as a mass functions. We applied it on a real-world biomedical opinion database. Performance analysis showed that our proposal generated better patterns compared to literature-based methods.

7.6. Mining Relevant Interval Rules

Participants: Philippe Besnard, Thomas Guyet, Véronique Masson, René Quiniou.

Rule mining is a classical data mining task. Numerical rule mining consists of extracting decision rules from a dataset with numerical attributes. In this work, we are interested in extracting a subset of accurate rules, called relevant rules. This selection criteria was introduced by Garriga et al. for categorical attributes [28]. In [13] we extend the method of Garriga et *al.* for mining relevant rules on numerical attributes by extracting interval-based pattern rules. We proposed an algorithm that extracts such rules from numerical datasets using the interval-pattern approach from Kaytoue et *al.* [29]. The algorithm has been implemented and intensively evaluated on real datasets. This study on numerical rules mining leads us to initiate a study about admissible generatizations of examples as rules [18].

7.7. Time Series Rule Matching: Application to Energy Consumption

Participants: Maël Guillemé, Véronique Masson, Laurence Rozé, René Quiniou, Alexandre Termier.

Pattern mining in time series is an important subfield of Data Mining. In various applications, patterns exhibit distortion in time (or time elasticity) that requires using specific distance measures. In this work, we extend an algorithm proposed by Shokoohi et *al.* [35] by improving the performance of rule matching in the detection of energy consumption patterns. Nowadays companies are more and more equipped with sensors in order to trace losses of energy resources. Detecting dysfunctions from time series recorded by these sensors becomes a crucial problem for reducing energy consumption. Locating specific patterns related to dysfunctions in time series requires handling with time elasticity (i.e. distortion in time) of patterns. We propose a detection of predictive rules based on several variations of Dynamic Time Warping (DTW) and show the superiority of subsequence DTW [11]. We study now multivariate time series classification to predict dysfunctions as soon as possible.

7.8. Negative Temporal Sequence Mining

Participants: Katerina Tsesmeli, Thomas Guyet, René Quiniou, Manel Boumghar [EDF R&D], Laurent Pierre [EDF R&D].

Temporal pattern mining is one of the important tasks in the data mining research field. It aims at extracting interesting sequences of occurring events from timestamped event sequences as well as their temporal constraints relating sequence events. Little research has focused on mining sequential patterns with non-occurring (negative) events, though they can bring much value and relevance to extracted patterns. In this context, we are interested in formalizing normal and undesirable situations, that can be defined in terms of negative temporal patterns. We proposed the NTGSP algorithm [17] that extracts frequent sequences with positive and negative events, as well as temporal information about the delay between these events. The method performance has been evaluated on synthetic sequences and on commercial data provided by EDF, a major french power distribution company.

7.9. TopPI: An efficient algorithm for item-centric mining

Participants: Vincent Leroy, Martin Kirchgessner, Alexandre Termier, Sihem Amer-Yahia.

In this paper [6], we introduce item-centric mining, a new semantics for mining long-tailed datasets. Our algorithm, TopPI, finds for each item its top-k most frequent closed itemsets. While most mining algorithms focus on the globally most frequent itemsets, TopPI guarantees that each item is represented in the results, regardless of its frequency in the database.

TopPI allows users to efficiently explore Web data, answering questions such as “what are the k most common sets of songs downloaded together with the ones of my favorite artist?”. When processing retail data consisting of 55 million supermarket receipts, TopPI finds the itemset “milk, puff pastry” that appears 10,315 times, but also “frangipane, puff pastry” and “nori seaweed, wasabi, sushi rice” that occur only 1120 and 163 times, respectively. Our experiments with analysts from the marketing department of our retail partner, demonstrate that item-centric mining discover valuable itemsets. We also show that TopPI can serve as a building-block to approximate complex itemset ranking measures such as the p-value.

Thanks to efficient enumeration and pruning strategies, TopPI avoids the search space explosion induced by mining low support itemsets. We show how TopPI can be parallelized on multi-core architectures and Hadoop clusters. Our experiments on datasets with different characteristics show the superiority of TopPI when compared to standard top-k solutions, and to Parallel FPGrowth, its closest competitor.

7.10. Declarative Sequential Pattern Mining of Care Pathways

Participants: Thomas Guyet, André Happe [Brest University Hospital/REPERES], Yann Dauxais.

Sequential pattern mining algorithms are widely used to explore care pathways database, but they generate a deluge of patterns, mostly redundant or non-informative. Clinicians need tools to express complex mining queries in order to generate less but more significant patterns. These algorithms are not versatile enough to answer complex clinician queries. This work [12] proposes to apply a declarative pattern mining approach based on the Answer Set Programming paradigm. It is exemplified by a pharmaco-epidemiological study investigating the possible association between hospitalization for seizure and antiepileptic drug switch from a French medico-administrative database.

7.11. Efficiency Analysis of ASP Encodings for Sequential Pattern Mining Tasks

Participants: Thomas Guyet, Yves Moinard, René Quiniou, Torsten Schaub.

This study [22] presents the use of Answer Set Programming (ASP) to mine sequential patterns. ASP is a high-level declarative logic programming paradigm that allows for representation of combinatorial and optimization problems, as well as knowledge and reasoning tasks. Thus, ASP is a good candidate for implementing pattern mining with background knowledge, which has been a data mining issue for a long time. We propose encodings of the classical sequential pattern mining tasks within two representations of embeddings (fill-gaps vs skip-gaps) and for various kinds of patterns: frequent, constrained and condensed. We compare the computational performance of these encodings with each other to get a good insight into the efficiency of ASP encodings. The results show that the fill-gaps strategy is better on real problems due to lower memory consumption. Finally, compared to a constraint programming approach (CPSM), another declarative programming paradigm, our proposal showed comparable performance.

7.12. Mining Rare Sequential Patterns with ASP

Participants: Ahmed Samet, Thomas Guyet, Benjamin Négrevergne.

This work [20] presents an approach of meaningful rare sequential pattern mining based on the declarative programming paradigm of Answer Set Programming (ASP). The setting of rare sequential pattern mining is introduced. Our ASP approach provides an easy manner to encode expert constraints on expected patterns to cope with the huge amount of meaningless rare patterns. Encodings are presented and quantitatively compared to a procedural baseline. An application on care pathways analysis illustrates the applicability of our method in the encoding of constraints provided by experts.

7.13. From Medico-administrative Databases Analysis to Care Trajectories Analytics: An Example with the French SNDS

Participants: Erwan Drezen [Rennes University Hospital/REPERES], Thomas Guyet, André Happe [Brest University Hospital/REPERES].

Medico-administrative data like SNDS (Système National de Données de Santé) are not collected initially for epidemiological purposes. Moreover, the data model and the tools proposed to SNDS users make their in-depth exploitation difficult. We propose a data model, called the ePEPS model, based on health care trajectories to provide a medical view of raw data [4]. A data abstraction process enables the clinician to have an intuitive medical view of raw data and to design a study-specific views. This view is based on a generic model of care trajectory, i.e. a sequence of timestamped medical events for a given patient. This model is combined with tools to manipulate care trajectories efficiently.

7.14. A Data Warehouse to Explore Multidimensional Simulated Data from a Spatially Distributed Agro-hydrological Model to Improve Catchment Nitrogen Management

Participants: Tassadit Bouadi, Marie-Odile Cordier, Pierre Moreau, Jordy Salmon-Monviola, Chantal Gascuel-Odoux.

Spatially distributed agro-hydrological models allow researchers and stakeholders to represent, understand and formulate hypotheses about the functioning of agro-environmental systems and to predict their evolution. These models have guided agricultural management by simulating effects of landscape structure, farming system changes and their spatial arrangement on stream water quality. Such models generate many intermediate results that should be managed, analyzed and transformed into usable information. We introduce [3] a data warehouse (N-Catch) built to store and analyze simulation data from the spatially distributed agro-hydrological model TNT2. We present scientific challenges to and tools for building data warehouses and describe the three dimensions of N-Catch: space, time and an original hierarchical description of cropping systems. We show how to use OLAP to explore and extract all kinds of useful high-level information by aggregating the data along these three dimensions. We also show how to facilitate exploration of the spatial dimension by coupling N-Catch with GIS. Such tool constitutes an efficient interface between science and society, simulation remaining a research activity, exploration of the results becoming an easy task accessible for a large audience.

7.15. Extended Automata for Temporal Planning of Interacting Agents

Participants: Christine Largouët, Omar Krichen, Yulong Zhao.

In this paper [5], we consider the planning problem for a system represented as a set of interacting agents evolving along time according to explicit timing constraints. Given a goal, the planning problem is to find the sequence of actions such that the system reaches the goal state in a limited time and in an optimal manner, assuming actions have a cost. In our approach, the planning problem is based on model-checking and controller synthesis techniques while the goal is defined using temporal logic. Each agent of the system is represented 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 the definition of uncontrollable actions. We define a planning algorithm that computes the best strategy to achieve a goal. To experiment our approach, we extend the classical Transport Domain with timing constraints, cost on actions and uncontrollable actions. The planning algorithm is finally presented on a marine ecosystem management problem.

7.16. Anomaly Detection in Streams with Extreme Value Theory

Participants: Alban Siffer [EMSEC], Pierre-Alain Fouque [EMSEC], Christine Largouët, Alexandre Termier.

Anomaly detection in time series has attracted considerable attention due to its importance in many real-world applications including intrusion detection, energy management and finance. Most approaches for detecting outliers rely on either manually set thresholds or assumptions on the distribution of data. In [16], we propose a new approach to detect outliers in streaming univariate time series based on Extreme Value Theory that does not require to handpick thresholds and makes no assumption on the distribution: the main parameter is only the risk, controlling the number of false positives. Our approach can be used for outlier detection, but more generally for automatically setting thresholds, making it useful in wide number of situations. We also test our algorithms on various real-world datasets which confirm the soundness and efficiency of our methods.

LAGADIC Project-Team

7. New Results

7.1. Visual Perception

7.1.1. Visual Tracking for Motion Capture

Participant: Eric Marchand.

This work is achieved in collaboration with Anatole Lécuyer (Inria Hybrid group) through the co-supervision of Guillaume Cortes Ph.D.

In the context of the development of new optical tracking devices, we propose an approach to greatly increase the tracking workspace of VR applications without adding new sensors [69]. Our approach relies on controlled cameras able to follow the tracked markers all around the VR workspace providing 6 DoF tracking data. We designed the proof-of-concept of such approach based on two consumer-grade cameras and a pan-tilt head. This approach has also been extended for the tracking of a drone in GPS denied environment [42].

We also achieved a short study related to the analysis of the 3D motion of head and hand in CAVE-based applications with the goal to optimize optical tracking sensors placement [43].

7.1.2. Object 3D Tracking based on Depth Information and CAD Model

Participants: Agniva Sengupta, Eric Marchand, Alexandre Krupa.

In the context of the iProcess project (see Section 9.3.3.2), we started this year a new study related to pose estimation and tracking of a rigid object observed by a RGB-D camera. We developed a pose estimation approach based on depth information measurement and the use of a CAD model represented by a 3D tetrahedral mesh. The pose parameters are estimated through an iterative optimization process that minimizes the point-to-plane Euclidean distance between the point cloud observed by the RGB-D camera and the surface of the 3D mesh. Preliminary results obtained with simple objects constituted by a set of orthogonal planes showed good performance of this approach. However, the method failed for the case of complex objects that exhibit important curvature surfaces. In order to address this issue we are currently extending the approach to take into account also the RGB information in the optimization criterion.

7.1.3. General Model-based Tracker

Participants: Souriya Trinh, Fabien Spindler, François Chaumette.

We have generalized the model-based tracker [2] available in ViSP [5] to integrate the depth information provided by a RGB-D sensor using the method described in the previous paragraph. It is now possible to fuse in the same optimization scheme measurements such as points of interest, edges, and depth, which allows to improve the robustness and accuracy of the tracker.

7.1.4. 3D Localization for Airplane Landing

Participants: Noël Mériaux, Pierre-Marie Kerzerho, Patrick Rives, Eric Marchand, François Chaumette.

This study was 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 implemented a direct registration method based on dense vision-based tracking that allows localizing the on-board camera from a set of keyframe images corresponding to the landing trajectory. First experiments with simulated and real images have been carried on with promising results. This approach is particularly interesting at the beginning of the descent when the landing track is far away and not very observable in the image. In that sense, the direct registration method is strongly complementary with the model-based approach studied before.

7.1.5. Extrinsic Calibration of Multiple RGB-D Cameras

Participants: Eduardo Fernandez Moral, Patrick Rives.

In collaboration with Alejandro Perez-Yus from the University of Zaragoza, we developed a novel method to estimate the relative poses between RGB and depth cameras without the requirement of an overlapping field of view, thus providing flexibility to calibrate a variety of sensor configurations. This calibration problem is relevant to robotic applications which can benefit of using several cameras to increase the field of view. In our approach, we extract and match lines of the scene in the RGB and depth cameras, and impose geometric constraints to find the relative poses between the sensors. In [31], an analysis of the observability properties of the problem is presented. We have validated our method in both synthetic and real scenarios with different camera configurations, demonstrating that our approach achieves good accuracy and is very simple to apply, in contrast with previous methods based on trajectory matching using visual odometry or SLAM.

7.1.6. Scene Registration with Large Convergence Domain

Participants: Renato José Martins, Patrick Rives.

Image registration has been a major problem in computer vision over the past decades. It implies searching an image in a database of previously acquired images to find one (or several) that fulfill some degree of similarity, e.g. an image of the same scene from a similar viewpoint. This problem is interesting in mobile robotics for topological mapping, re-localization, loop closure and object identification. Scene registration can be seen as a generalization of the above problem where the representation to match is not necessarily defined by a single image (i.e. the information may come from different images and/or sensors), attempting to exploit all information available to pursue higher performance and flexibility. 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 [12]. The main idea is to explore the properties given by planar surfaces with co-visibility and their normals from two distinct viewpoints. We estimate, in two decoupled stages, the rotation and then the translation, both based on the normal vectors orientation and on the depth. These two stages are efficiently computed by using low resolution depth images and without any feature extraction/matching. In [53], we also analyze the limitations and observabilty of this approach, and its relationship to ICP point-to-plane. Notably, if the rotation is observable, at least five DoF can be estimated in the worst case. To demonstrate the effectiveness of the method, we evaluate the initialization technique in a set of challenging scenarios, comprising simulated spherical images from the Sponza Atrium model benchmark and real spherical indoor sequences.

7.1.7. Scene Semantization based on Deep Learning Approach

Participants: Eduardo Fernandez Moral, Patrick Rives.

Semantic segmentation of images is an important problem for mobile robotics and autonomous driving because it offers basic information which can be used for complex reasoning and safe navigation. This problem constitutes a very active field of research, where the state-of-the-art evolves continuously with new strategies based on different kinds of deep neural networks for image segmentation and classification. RGB-D images are starting to be employed as well for the same purpose to exploit complimentary information from color and geometry. The team LAGADIC has explored several strategies to increase the performance and the accuracy of semantic segmentation from RGB-D images. We propose a multi-pipeline architecture to exploit effectively the complimentary information from RGB-D images and thus to improve the semantic segmentation results. The multi-pipeline architecture processes the color and depth layers in parallel, before concatenating their feature maps to produce the final semantic prediction. Our results are evaluated on public benchmark datasets to show the improved accuracy of the proposed architecture. [46] Though we address this problem in the context of urban images segmentation, our results can also be extended to other contexts, like indoor scenarios and domestic robotics.

Our research is partly motivated by the need of semantic segmentation solutions with better segmentation around contours. Besides, we note that one of the main issues when comparing different neural networks architectures is how to select an appropriate metric to evaluate their accuracy. We have studied several metrics for multi-class classification, and we propose a new metric which accounts for both global and contour accuracy in a simple formulation to overcome the weaknesses of previous metrics. This metric is based on the Jaccard index, and takes explicitly into account the distance to the border regions of the different classes, to encode jointly the rate of correctly labeled pixels and how homeomorphic is the segmentation to the real object boundaries. We also present a comparative analysis of our proposed metric and several commonly used metrics for semantic segmentation together with a statistical analysis of their correlation.

7.1.8. Online Localization and Mapping for UAVs

Participants: Muhammad Usman, Paolo Robuffo Giordano.

Localization and mapping in unknown environments is still an open problem, in particular for what concerns UAVs because of the typical limited memory and processing power available onboard. In order to provide our quadrotor UAVs with high autonomy, we started studying how to exploit onboard cameras for an accurate (but fast) localization and mapping in unknown indoor environments. We chose to base both processes on the newly available Semi-Direct Visual Odometry (SVO) library (http://rpg.ifi.uzh.ch/software) which has gained considerable attention over the last years in the robotics community. The idea is to exploit dense images (i.e., with little image pre-processing) for obtaining an incremental update of the camera pose which, when integrated over time, can provide the camera localization (pose) w.r.t. the initial frame. In order to reduce drifts during motion, a concurrent mapping thread is also used for comparing the current view with a set of keyframes (taken at regular steps during motion) which constitute a "map" of the environment. We have started porting the SVO library to our UAVs and the preliminary results showed good performance of the localization accuracy against the Vicon ground truth. We are now planning to close the loop and base the UAV flight on the reconstructed pose from the SVO algorithm.

7.1.9. Reflectance and Illumination Estimation for Realistic Augmented Reality

Participants: Salma Jiddi, Eric Marchand.

A key factor for realistic Augmented Reality is a correct illumination simulation. This consists in estimating the characteristics of real light sources and use them to model virtual lighting. This year, we studied a novel method for recovering both 3D position and intensity of multiple light sources using detected cast shadows. Our algorithm has been successfully tested on a set of real scenes where virtual objects have visually coherent shadows [70].

7.1.10. Optimal Active Sensing Control

Participants: Marco Cognetti, Paolo Salaris, Paolo Robuffo Giordano.

This study concerns the problem of active sensing control whose objective is to reduce the estimation uncertainty of an observer as much as possible by determining the inputs of the system that maximize the amount of information gathered by the few noisy outputs while at the same time reduce the negative effects of the process/actuation noise. The latter is far from being negligible for several robotic applications (a prominent example being aerial vehicles).

Last year, we extended a previous work [9] to the case where the observability property is not instantaneously guaranteed, and hence the optimal estimation strategy cannot be given in terms of the instantaneous velocity direction of the robot and consequently of the onboard sensors. These outcomes of this research have been presented in [61] for nonlinear differentially flat systems. This year, we have moved some steps forward in order to improve and generalize the work in [61]: first of all, we have replaced the Observability Gramian (OG) with the Constructibility Gramian (CG). Despite their similar form, they differ from the fact that the OG measures the information collected along the path about the initial state of the nonlinear system while the CG measures the one about the current/final state with which most robotics applications are more concerned. Second, we have overcome the limit of previous work [61] that only deals with the case where the OG and

the CG are known in closed-form. We have also applied our method to the unicycle vehicle which is a more complex dynamic system than the one used in [61] and tested our machinery to the cases of self-calibration and environment reconstruction. Moreover, thanks to the arrival of Marco Cognetti in our group as Post-doc, we are currently working on the application of our method to a quadrotor UAV, which is a much more complex dynamic system, for which the CG is not known in closed-form. The ultimate goal is to test our new machinery in a real experiment with a quadrotor UAV. Finally, we have also worked on the problem of considering the process/actuation noise in the optimization algorithm. As the CG (or the OG) does not take into account the degrading effects on the information collected through the outputs of the process/actuation noise, we have proposed to directly maximize the smallest eigenvalue of the covariance matrix given by the Riccati differential equation of the EKF, used as estimation algorithm. The results of this approach have been submitted to ICRA 2018.

7.2. Sensor-based Robot Control

7.2.1. Determining Singularity Configurations in IBVS

Participant: François Chaumette.

This theoretical study has been achieved through an informal collaboration with Sébastien Briot and Philippe Martinet from LS2N in Nantes, France. It concerned the determination of the singularity configurations of image-based visual servoing using tools from the mechanical engineering community and the concept of "hidden" robot. In a first step, we have revisited the well-known case of using three image points as visual feature, and then solved the general case of n image points [16]. The case of three image straight lines has also been solved for the first time [17].

We have also designed a control scheme in order to avoid these singularities during the execution of a visual servoing scheme [38].

7.2.2. Visual Servoing through Mirror Reflection

Participants: François Chaumette, Eric Marchand.

Apart the use of catadioptric cameras, only few visual servoing works exploit the use of mirror. Such a configuration is however interesting since it allows overpassing the limited camera field of view. Based on the known projection equations involved in such a system, we studied the theoretical background that allows the control of planar mirror for visual servoing in different configurations. Limitations intrinsic to such systems, such as the number of DoF actually controllable, have been studied. The case of point feature was considered in [51] and this has been extended to line in [52].

7.2.3. Visual Servoing of Humanoid Robots

Participants: Giovanni Claudio, Fabien Spindler, François Chaumette.

This study is realized in the scope of the BPI Romeo 2 and H2020 Comanoid projects (see Sections 9.2.7 and 9.3.1.2).

We have designed the modeling of the visual features at the acceleration level to embed visual tasks and visual constraints in an existing Quadratic Programming controller [13]. Experimental results have been obtained on Romeo (see Section 6.8.4).

7.2.4. Model Predictive Visual Servoing

Participants: Paolo Robuffo Giordano, François Chaumette.

This study was realized in collaboration with Pierre-Brice Wieber, from Bipop group at Inria Rhône Alpes, through the co-supervision of Nicolas Cazy's Ph.D.

Model Predictive Control (MPC) is a powerful control framework able to take explicitly into account the presence of constraints in the controlled system (e.g., actuator saturations, sensor limitations, and so on). In this study, we studied the possibility of using MPC for tackling one of the most classical constraints of visual servoing applications, that is, the possibility to lose tracking of features because of occlusions, limited camera field of view, or imperfect image processing/tracking. The MPC framework depends upon the possibility to predict the future evolution of the controlled system over some time horizon, for correcting the current state of the modeled system whenever new information (e.g., new measurements) become available. We have also explored the possibility of applying these ideas in a multi-robot collaboration scenario where a UAV with a downfacing camera (with limited field of view) needs to provide localization services to a team of ground robots [41].

7.2.5. Model Predictive Control for Visual Servoing of a UAV

Participants: Bryan Penin, François Chaumette, Paolo Robuffo Giordano.

Visual servoing is a well-known class of techniques meant to control the pose of a robot from visual input by considering an error function directly defined in the image (sensor) space. These techniques are particularly appealing since they do not require, in general, a full state reconstruction, thus granting more robustness and lower computational loads. However, because of the quadrotor underactuation and inherent sensor limitations (mainly limited camera field of view), extending the classical visual servoing framework to the quadrotor flight control is not straightforward. For instance, for realizing a horizontal displacement the quadrotor needs to tilt in the desired direction. This tilting, however, will cause any downlooking camera to point in the opposite direction with, e.g., possible loss of feature tracking because of the limited camera field of view.

In order to cope with these difficulties and achieve a high-performance visual servoing of quadrotor UAVs, we chose to rely on MPC for explicitly dealing with this kind of constraints during flight. We have recently considered the problem of controlling in minimum-time a quadrotor UAV equipped with a downlooking camera that needs to reach a desired pose w.r.t. a target on the ground from visual input. The control problem is solved by an online replanning strategy that is able to generate (at camera rate) minimum-time trajectories towards the final pose while coping with actuation constraints (limited propeller thrusts) and sensing constraints (target always in the camera fov). By exploiting the camera images during motion, the replanning strategy is able to adjust online the optimal trajectory and, thus, be robust against unmodeled effects and other disturbances (which can be typically expected on a quadrotor flying aggressively). The approach has been validated via numerical simulations in [59]. We are now working towards an experimental validation, as well as novel algorithmic extensions allowing for the possibility of temporarily losing sight of the target object for relaxing the visibility constraint (and, thus, gain in maneuverability).

7.2.6. UAVs in Physical Interaction with the Environment

Participants: Quentin Delamare, Paolo Robuffo Giordano.

Most research in UAVs deals with either contact-free cases (the UAVs must avoid any contact with the environment), or in "static" contact cases (the UAVs need to exert some forces on the environment in quasi-static conditions, reminiscent of what has been done with manipulator arms). Inspired by the vast literature on robot locomotion (from, e.g., the humanoid community), in this research topic we aim at exploiting the contact with the environment for *helping* a UAV maneuvering in the environment, in the same spirit in which we humans (and, supposedly, humanoid robots) use our legs and arms when navigating in cluttered environments for helping in keeping balance, or perform maneuvers that would be, otherwise, impossible.

As an initial case study, we have considered a planar UAV equipped with a 1 DoF actuated arm capable of hooking at some pivots in the environment. This UAV (named MonkeyRotor) needs to "jump" from one pivot to the next one by exploiting the forces exchanged with the environment (the pivot) and its own actuation system (the propellers). This study considers the full dynamics in both cases (hooked, free-flying), proposes an optimization problem for finding optimal trajectories from an initial hooked configuration to the next one, and validates the approach in simulation. We are now working towards a physical realization of a first prototype. This activity is done in cooperation with LAAS-CNRS (Dr. Antonio Franchi who is co-supervising Quentin Delamare).

7.2.7. Visual Servoing for Steering Simulation Agents

Participants: Axel Lopez Gandia, Eric Marchand, François Chaumette, Julien Pettré.

Steering is one of the basic functionality of any character animation system. It provides characters with the ability to locally move in the environment so as to achieve basic navigation tasks, such as reaching a goal, avoiding a collision with an obstacles, etc. This problem has been explored in various contexts (e.g., motion planning, autonomous characters or crowd simulation). It turned out that this component plays an important role on the quality of character animation and received a lot of attention. Many important steps have been taken to improve steering techniques: potential fields, sets of attractive and repulsive forces, linear programming in the velocity space, local optimization of navigation functions, etc. Each new category of approach leads to characters close to forming realistic trajectories when achieving navigation.

Nevertheless, all these techniques remain quite far from the way real humans form their locomotion trajectory, because they are all based on kinematics and geometry. Humans obviously do not solve geometrical problems of this nature while moving in their environment but control their motion from perceptual features, and more especially visual features they perceive from the environment. For simulating more accurately the perception-action loop used by humans to navigate in their environment, we developed a technique which provides characters with vision capabilities, by equipping them with a virtual retina on which we project information about their surroundings. In a first version, we projected information about the relative motion of objects around them, allowing characters to estimate the risk of collision they face, and to move so as to minimize this risk [21]. More recently, we projected a purely visual information, and we established the relations that exist between the visual features characters perceive and the motion they perform. This way, we are able to steer characters so as their visual flow satisfies some conditions, allowing them for example to reach a goal while avoiding surrounding obstacles, could they be static or moving.

7.2.8. Direct Visual Servoing

Participants: Quentin Bateux, Eric Marchand.

We have proposed a deep neural network-based method to perform high-precision, robust and real-time 6 DoF visual servoing [63]. We studied how to create a dataset simulating various perturbations (occlusions and lighting conditions) from a single real-world image of the scene. A convolutional neural network is fine-tuned using this dataset to estimate the relative pose between two images of the same scene. The output of the network is then employed in a visual servoing control scheme. The method converges robustly even in difficult real-world settings with strong lighting variations and occlusions.

7.3. Medical Robotics

7.3.1. Visual Servoing using Wavelet and Shearlet Transforms

Participants: Lesley-Ann Duflot, Alexandre Krupa.

In collaboration with Femto-ST lab in Besançon and the Research Group on Computational Data Analysis at Universitat Bremen, we developed a new generation of direct visual servoing methods in which the signal control inputs are the coefficients of a multiscale image representation. In particular, we considered the use of multiscale image representations that are based on discrete wavelet and shearlet transforms. We succeeded to derive an analytical formulation of the interaction matrix related to the wavelet and shearlet coefficients and experimentally demonstrated the performances of the proposed visual servoing approaches. We also considered this control framework in the design of a medical application which consists in automatically moving a biological sample carried by a parallel micro-robotic platform using Optical Coherence Tomography (OCT) as visual feedback. The objective is to automatically retrieve the region of the sample that corresponds to an initial optical biopsy for diagnosis purpose. First results obtained with a 3 DoF eye-to-hand visual servoing demonstrated the feasibility to use the wavelet coefficients of the OCT image as input of the control law.

7.3.2. 3D Steering of Flexible Needle by Ultrasound Visual Servoing

Participants: Jason Chevrie, Marie Babel, Alexandre Krupa.

We pursued our work on 3D steering of a flexible needle using ultrasound visual servoing [11]. This year, in collaboration with the Surgical Robotics Laboratory of the University of Twente, we developed a method to control a 2 DoF needle insertion device attached to the end-effector of a 6-DoF robotic arm in order to automatically insert a flexible needle toward a spherical target embedded in a moving biological tissue (bovine liver). We proposed a method that uses both base manipulation control and tip-based control while compensating the tissue motion to avoid lateral tearing. The visual feedback provided by the ultrasound probe was used to track the target and an electromagnetic tracker attached inside the needle was used to locate its tip. In this study, the motion compensation of the moving tissue was performed by minimizing the interaction force measured at the base of the needle insertion device. In our approach we used the generic task control framework to fuse the needle targeting and motion compensation tasks into a single control law. First experimental ex-vivo results demonstrated the efficiency of the proposed control to reach a target in moving biological tissue.

7.3.3. Robotic Assistance for Ultrasound Elastography by Visual Servoing, Force Control and Teleoperation

Participants: Pedro Alfonso Patlan Rosales, Alexandre Krupa.

This work concerns the development of a robotic assistant system for quantitative ultrasound elastography. This imaging modality provides the elastic parameters of a tissue which are commonly related with a certain pathology. It is performed by applying continuous stress variation on the tissue in order to estimate a strain map. Usually, this stress variation is performed manually by the user through the manipulation of the ultrasound probe and it results therefore in an user-dependent quality of the strain map. To improve the ultrasound elastography imaging and provide quantitative measurement, we developed an assistant robotic palpation system that automatically moves a 2D ultrasound probe for optimizing ultrasound elastography [72]. This year we extended our previous robotic palpation system in order to perform 3D elastography and allow the user to teleoperate the probe orientation through a haptic device [56]. This extension is based on the use of a 3D ultrasound probe held by a 6 DoF robotic arm and the design of a new control law based on the task control framework that simultaneously performs three tasks: i) autonomous palpation by force control of the tissue required for the strain map estimation, ii) probe lateral alignment on a stiff target of interest for optimizing its visibility by visual servoing and iii) teleoperation of the probe orientation by the user for exploration purpose. Recently, we also proposed a solution that allows the estimation of the strain map of a moving tissue that is subject to physiological motion [57]. It is based on the combination of a non-rigid motion tracking of the tissue of interest in the ultrasound image and an automatic 6 DoF compensation of the perturbation motion by visual servoing using dense ultrasound information.

7.3.4. Haptic Guidance of a Biopsy Needle

Participants: Hadrien Gurnel, Alexandre Krupa.

We started a new study in collaboration with Maud Marchal (Inria Hybrid group) related to the assistance of manual needle steering for biopsies or therapy purposes (see Section 9.1.7). Instead of automatically inserting the needle by a robotic arm as we did in other works, our objective is to develop a solution that provides haptic cue feedback to the clinician that helps him during its manual gesture. The haptic cue feedback will be provided by a haptic device holding the needle. This year we developed a software tool that simulates and visualizes the interaction of a virtual needle with a deformable virtual organ. This organ is represented by a 3D mesh and a mass-spring-damper model was considered to simulate its deformation due to the needle insertion motion. The development of this software was based on our libraries UsTk and ViSP and the external library VTK (Visualization Toolkit). We also interfaced to this simulator our Haption Virtuose 6D haptic device to allow the user to teleoperate the virtual needle and to feel the force applied by the needle on the virtual tissue. This simulator will constitute an important tool for our future development of dynamic haptic guides before testing them in a real experimental setup.

7.4. Teleoperation

7.4.1. Shared Control for Remote Manipulation

Participants: Firas Abi Farraj, Paolo Robuffo Giordano.

This work concerns our activities in the context of the RoMaNS H2020 project (see Section 9.3.1.3). Our main goal is to allow a human operator to be interfaced in an intuitive way with a two-arm system, one arm carrying a gripper (for grasping an object), and the other one carrying a camera for looking at the scene (gripper + object) and providing the needed visual feedback. The operator should be allowed to control the two-arm system in an easy way for letting the gripper approaching the target object, and she/he should also receive force cues informative of how feasible her/his commands are w.r.t. the constraints of the system (e.g., joint limits, singularities, limited camera fov, and so on).

We have started working on this topic by proposing a shared control architecture in which the operator could provide instantaneous velocity commands along four suitable task-space directions not interfering with the main task of keeping the gripper aligned towards the target object (this main task was automatically regulated). The operator was also receiving force cues informative of how much her/his commands were conflicting with the system constraints, in our case joint limits of both manipulators. Finally, the camera was always moving so as to keep both the gripper and the target object at two fixed locations on the image plane. Recently, we have extended this framework in several directions:

1. in a first extension, the existing instantaneous interface has been improved towards an "integral" approach in which the user can command parts of the future manipulator trajectory, while the autonomy makes sure that no constraint is violated (in this case we considered, again, joint limits and singularities, as well as a more realistic vision constraint for keeping the gripper and the object always in visibility and not overlapping). This shared control algorithm was validated in simulation in [58]. We are currently completing a full implementation on our dual-arm system (the two Viper robots);
2. second, we have studied how to integrate learning from demonstration into our framework by first using learning techniques for extracting statistical regularities of "expert users" executing successful trajectories for the gripper towards the target object. Then, these learned trajectories were used for generating force cues able to guide novice users during their teleoperation task by the "hands" of the expert users who demonstrated the trajectories in the first place [37];
3. third, we have considered a grasping scenario in which a post-grasp task is specified (e.g., the grasped object needs to follow a predefined trajectory): in this scenario, the operator (supported by the robot autonomy) needs to decide where to best grasp in order to then execute the desired post-grasp action. However, different grasping poses will result in easier/harder execution by the robot because of any possible constraint (e.g., joint limits and singularities). Since awareness of these constraints is hard for any operator, in this case the autonomy component cues the operator with a force feedback indicating the best grasp pose w.r.t. the existing constraints and post-grasp task. The operator has still control over where to grasp, but she/he is guided by the force feedback into more feasible grasp poses than what she/he could have guessed without any feedback [48];
4. finally, we have considered the task of assisting an operator in control of a UAV which is mapping a remote environment with an onboard camera. In this scenario the operator can control the UAV motion during the mapping task. However, as in any estimation problem, different motions will result less/more optimal w.r.t. the scene estimation task: therefore, a force feedback is produced in order to assist the operator in selecting the UAV motion (in particular, its linear velocity) that also results optimal for the sake of facilitating the scene estimation process. The results have been validated with numerical simulations in a realistic environment [39].

7.4.2. Wearable haptics

Participants: Marco Aggravi, Claudio Pacchierotti.

Kinesthetic haptic feedback is used in robotic teleoperation to provide the human operator with force information about the status of the slave robots and their interaction with the remote environment. Although kineshetic feedback has been proven to enhance the performance of teleoperation systems, it still shows several limitations, including its negative effect on the safety and stability of such systems, or the limited workspace, available DoF, high cost, and complexity of kinesthetic interfaces. In this respect, wearable haptics is gaining great attention. Safe, compact, unobtrusive, inexpensive, easy-to-wear, and lightweight haptic devices enable researchers to provide compelling touch sensations to multiple parts of the body, significantly increasing the applicability of haptics in many fields, such as robotics, rehabilitation, gaming, and immersive systems.

In this respect, our objective has been to study, design, and evaluate novel wearable haptic interfaces for the control of remote robotic systems as well as interacting with virtual immersive environments.

We have started by working on a multi-point wearable feedback solution for robotic manipulators operating in a cluttered environment [40]. The slave system is composed of an anthropomorphic soft robotic hand attached to a 6-axis force-torque sensor, which is in turn fixed to a 6-DoF robotic arm. The master system is composed of a Leap Motion controller and two wearable vibrotactile armbands, worn on the forearm and upper arm. The Leap Motion tracks the user's hand pose to control the pose of the manipulator and the grasping configuration of the robotic hand. The armband on the forearm conveys information about collisions of the slave hand/wrist system (green patch to green armband, see Fig. 8), whereas the armband on the upper arm conveys information about collisions of the slave arm (orange patch to orange armband). The amplitude of the vibrotactile feedback relayed by the armbands is proportional to the interaction force of the collision. A camera mounted near the manipulator's end-effector enables the operator to see the environment in front of the robotic hand. To validate our system, we carried out a human subjects telemanipulation experiment in a cluttered scenario. Twelve participants were asked to control the motion of the robotic manipulator to grasp an object hidden between debris of various shapes and stiffnesses. Haptic feedback provided by our wearable devices significantly improved the performance of the considered telemanipulation tasks. Finally, all subjects but one preferred conditions with wearable haptic feedback.

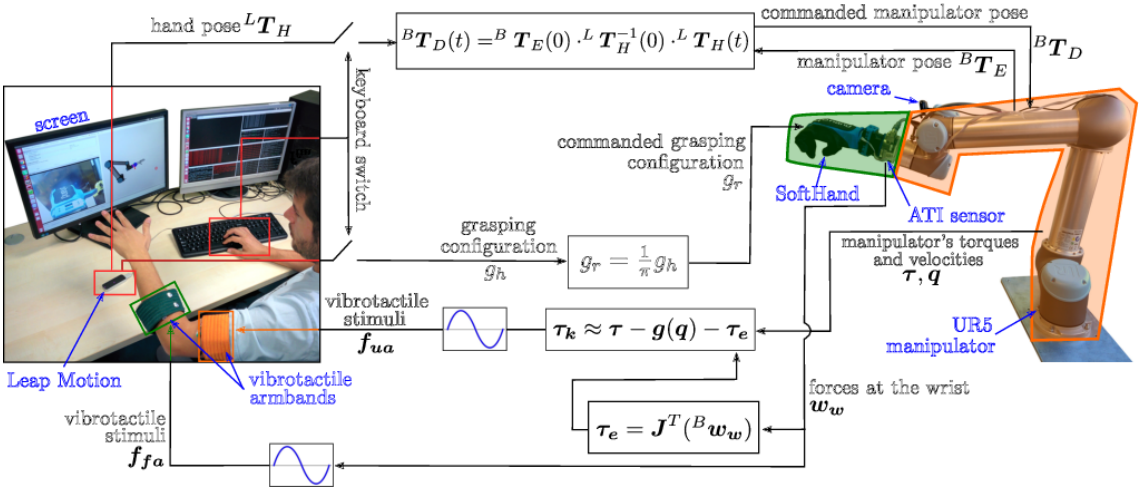

Figure 8. Haptic-enabled teleoperation system. We used two vibrotactile wearable devices to provide multi-point haptic feedback about collisions of the slave robot with the remote environment [40].

We have also used wearable haptics for guidance [20]. In this context, haptic feedback is not used to provide information about a force exerted by the salve robot in the remote environment, but it provides guidance cues about a predetermined trajectory to follow. Toward this, we developed a novel wearable device for the

forearm. Four cylindrical rotating end effectors, located on the user's forearm, can generate skin stretch at the ulnar, radial, palmar, and dorsal sides of the arm. When all the end effectors rotate in the same direction, the cutaneous device is able to provide cues about a desired pronation/supination of the forearm. On the other hand, when two opposite end effectors rotate in opposite directions, the device is able to provide cutaneous cues about a desired translation of the forearm. Combining these two stimuli, we can provide both rotation and translation guidance. To evaluate the effectiveness of our device in providing navigation information, we carried out two experiments of haptic navigation. In the first one, subjects were asked to translate and rotate the forearm toward a target position and orientation, respectively. In the second experiment, subjects were asked to control a 6-DoF robotic manipulator to grasp and lift a target object. Haptic feedback provided by our wearable device improved the performance of both experiments with respect to providing no haptic feedback. Moreover, it showed similar performance with respect to sensory substitution via visual feedback, without overloading the visual channel.

Finally, we also used wearable haptics for immersive virtual and augmented reality experiences, mainly addressing tasks related to entertainment and industrial training. In these case, we used wearable devices for the fingertips able to provide pressure and skin stretch sensations [24]. This article has also been featured in the News section of Science Magazine.

We also presented a review paper on the topic of wearable haptic devices for the hand [29].

7.5. Navigation of Mobile Robots

7.5.1. Visual Navigation from an Image Memory

Participants: Paolo Robuffo Giordano, François Chaumette.

This study achieved during Suman Raj Bista's Ph.D. was 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 localization as well as for the visual servoing are not composed of points of interest only, but on a combination of points of interest and straight lines since they are more common indoors [60]. Satisfactory experimental results have been obtained using the Pioneer mobile robot (see Section 6.8.2) and Pepper (See Section 6.8.4).

7.5.2. Robot-Human Interactions during Locomotion

Participant: Julien Pettré.

In collaboration with the Gepetto team of Laas in Toulouse and the Mimetic group in Rennes, we have studied how humans avoid collision with a robot. Understanding how humans achieve such avoidance is crucial to better anticipate humans' reactions to the presence of a robot and to control the robot to adapt its trajectory accordingly. It is generally assumed that humans avoid a robot just like they avoid another human. Last year, we brought the empirical evidence that humans actually set a specific strategy to avoid robots: they showed a preference to give way to a robot [36]. However, the robot was passive, i.e., not reacting to the presence of participants. This year, we studied interactions between humans and reactive robot, performing avoidance maneuvers to avoid collisions. Our conclusions are that, in such situations of human-robot interactions, human behave again as during human-human avoidance interactions. Again, this study provides useful guidelines about the design of robot control techniques.

7.5.3. Semi-Autonomous Control of a Wheelchair for Navigation Assistance

Participants: Louise Devigne, Marie Babel.

In order to improve the access to mobility for people with disabilities, we have previously designed a semi-autonomous assistive wheelchair system which progressively corrects the trajectory as the user manually drives the wheelchair and smoothly avoids obstacles. Within the frame of ISI4NAVE associated team (see Section 9.4.1.2), we investigated probabilistic blending approaches which take into account uncertainty in the interaction [45]. We also designed a shared-control curb-following solution for outdoor assisted power wheelchair navigation. Once a curb is detected, user input is blended with constraints deduced from the distance from sensors to the detected curb. This provides an intuitive shared control scheme capable of assisting the user while needed i.e. while approaching a curb. Preliminary validation tests of the robotic system were conducted within the PAMELA facility.

Developing and testing such systems for wheelchair driving assistance requires a significant amount of material resources and clinician time. With Virtual Reality technology, prototypes can be developed and tested in a risk-free and highly flexible Virtual Environment before equipping and testing a physical prototype. Additionally, users can "virtually" test and train more easily during the development process. We then designed a power wheelchair driving simulator allowing the user to navigate with a standard wheelchair in an immersive 3D Virtual Environment. In order to validate the framework including the driving assistance solution, we performed tests on the Immersia platform (Inria Hybrid team) with able-bodied participants and we have shown that the simulator it generates a good sense of presence and requires rather low cognitive effort from users [44].

7.5.4. Wheelchair Kinematics and Dynamics Modeling for Shared Control

Participants: Aline Baudry, Marie Babel.

The driving experience of an electric powered wheelchair can be disturbed by unpleasant dynamic effects of the caster wheels, particularly during maneuvers in narrow rooms and direction changes. In order to prevent their nasty behaviour, we propose to model caster wheel kinematics and dynamics in order to implement a control law for a semi-autonomous assistance to maneuver in narrow environments. We conducted a preliminary study that has been achieved for our three types of wheelchair, each presenting different kinematic behaviors: front caster type, rear caster type and mid-wheel drive (see Figure 3 .c). Transfer functions for each of these configurations have been identified. We achieved to design a parametric transfer function of the caster's behavior regarding to the initial orientation, wheelchair's velocity and user mass, in order to develop a sensorless maneuver control law.

7.5.5. Wheelchair Autonomous Navigation for Fall Prevention

Participants: Solenne Fortun, Marie Babel.

The Prisme project (see Section 9.1.8) is devoted to fall prevention and detection of inpatients with disabilities. For wheelchair users, falls typically occur during transfer between the bed and the wheelchair and are mainly due to a bad positioning of the wheelchair. In this context, the Prisme project addresses both fall prevention and detection issues by means of a collaborative sensing framework. Ultrasonic sensors are embedded onto both a robotized wheelchair and a medical bed. The measured signals are used to detect fall and to automatically drive the wheelchair near the bed at an optimal position determined by occupational therapists. We first designed a detection solution based on a multiple echoes technique that enhances the system perception abilities. This augmented perception system is planned to be used for wheelchair navigation as well as fall detection.

7.5.6. Robotic Platform for Assistance to People with Reduce Mobility

Participants: Dayana Hassan, Paolo Salaris, Patrick Rives.

The main objective of this work is to develop, in collaboration with AXYN Robotics (see Section 8.2.4), an intelligent vehicle to help elderly or persons with reduced mobility to move safely within a retirement home, an hospital or other much more crowded and dynamic environments. First of all, the vehicle has to be able to move within the environment while at the same time update the current map as accurately as possible. Once the map of the environment is available, the robot has to be able to plan the trajectory and reach a given destination. The robot should also follow a person taking into account social behaviors or bring towards a

given destination, e.g. the canteen, making sure that an elderly person, affected e.g. by Alzheimer's disease, follows the robot. The robot should also work as an intelligent walker and help people in case of falling. In all these cases, it is very important to include humans (i.e. his/her model, his/her behaviors, his/her intentions etc.) within the study in order to develop adaptable human-aware path planning and control strategies. During this first year, the problem of following a person has been studied, starting from the literature, in order to find a suitable control scheme that merges feedback control laws, aimed at reactively cope with neighborhood environment events and feedforward ones, mainly intended to take into account the intentions of the person to follow, also including social behaviors.

7.6. Multi-robot and Crowd Motion Control

7.6.1. Rigidity-based Methods for Formation Control

Participants: Fabrizio Schiano, Paolo Robuffo Giordano.

Most multi-robot applications must rely on *relative sensing* among the robot pairs (rather than absolute/external sensing such as, e.g., GPS). 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 our previous works we have addressed the problem of coordinating a team of quadrotor UAVs equipped with onboard cameras from which one could extract "relative bearings" (unit vectors in 3D) w.r.t. the neighboring UAVs in visibility. This problem is known as bearing-based formation control and localization. The basic assumption, however, was to always have a bearing rigid graph which may easily conflict with any sensing/communication constraint (measurements/edges can be lost whenever, e.g., a UAV leaves the camera fov, or it is occluded by another UAV/obstacle). In [62] we have then tackled the problem of "bearing rigidity maintenance" by studying how to formalize the problem of maintaining bearing rigidity over time despite possible sensing/communication constraints (min/max range, limited camera fov and occlusions in the reported work). Thanks to a suitable weighing machinery, we could define a "bearing rigidity eigenvalue" as a suitable metric for quantifying the degree of rigidity in the interaction graph, and then we could propose a gradient-based controller able to maintain the rigidity eigenvalue always positive (and, thus, guarantee bearing rigidity maintenance). The approach has been validated by experiments run on 5 quadrotor UAVs.

7.6.2. Cooperative Localization using Interval Analysis

Participants: Ide Flore Kenmogne Fokam, Vincent Drevelle.

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 bounded-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. Results have been obtained for image-based localization of a single UAV, enabling to characterize the pose uncertainty domain from measurements uncertainties [50], and also fusion with onboard proprioceptive sensors [49]. Extension to cooperative localization in a multi-UAV fleet has been studied in the two-robot case and continues as an ongoing work.

LINKMEDIA Project-Team

7. New Results

7.1. Multimedia indexing, Motif and knowledge discovery

7.1.1. Towards engineering a web-scale multimedia service: a case study using SPARK

Participant: Laurent Amsaleg.

Joint work with Gylfi Þór Guðmundsson (Univ. Reykyavik), Björn Þór Jónsson (Univ. Copenhagen) and Michael J. Franklin (UC Berkeley).

Computing power has now become abundant with multi-core machines, grids and clouds, but it remains a challenge to harness the available power and move towards gracefully handling web-scale datasets. Several researchers have used automatically distributed computing frameworks, notably Hadoop and Spark, for processing multimedia material, but mostly using small collections on small clusters. We describe the engineering process for a prototype of a (near) web-scale multimedia service using the Spark framework running on the AWS cloud service. We present experimental results using up to 43 billion SIFT feature vectors from the public YFCC 100M collection, making this the largest high-dimensional feature vector collection reported in the literature. The design of the prototype and performance results demonstrate both the flexibility and scalability of the Spark framework for implementing multimedia services.

7.1.2. On competitiveness of nearest-neighbor based music classification: a methodological critique

Participant: Laurent Amsaleg.

Joint work with Haukur Pálmasson, Björn Þór Jónsson (Univ. Copenhagen), Markus Schedl (Johannes Kepler University), Peter Knees (TU Wien).

The traditional role of nearest-neighbor classification in music classification research is that of a straw man opponent for the learning approach of the hour. Recent work in high-dimensional indexing has shown that approximate nearest-neighbor algorithms are extremely scalable, yielding results of reasonable quality from billions of high-dimensional features. With such efficient large-scale classifiers, the traditional music classification methodology of reducing both feature dimensionality and feature quantity is incorrect; instead the approximate nearest-neighbor classifier should be given an extensive data collection to work with. We present a case study, using a well-known MIR classification benchmark with well-known music features, which shows that a simple nearest-neighbor classifier performs very competitively when given ample data. In this position paper, we therefore argue that nearest-neighbor classification has been treated unfairly in the literature and may be much more competitive than previously thought [30].

7.1.3. Unsupervised part learning for visual recognition

Participants: Ronan Sicre, Yannis Avrithis, Ewa Kijak.

Joint work with Frederic Jurie (Univ. Caen).

Part-based image classification aims at representing categories by small sets of learned discriminative parts, upon which an image representation is built. Considered as a promising avenue a decade ago, this direction has been neglected since the advent of deep neural networks. In this context, the work proposed here brings two contributions: first, this work proceeds one step further compared to recent part-based models (PBM), focusing on how to learn parts without using any labeled data. Instead of learning a set of parts per class, as generally performed in the PBM literature, the proposed approach constructs a partition of a given set of images into visually similar groups, and subsequently learns a set of discriminative parts per group in a fully unsupervised fashion. This strategy opens the door to the use of PBM in new applications where labeled data are typically not available, such as instance-based image retrieval. Second , we show that despite the recent success of end-to-end models, explicit part learning can still boost classification performance. We experimentally show that our learned parts can help building efficient image representations, which outperform state-of-the art deep convolutional neural networks on both classification and retrieval tasks [32].

7.1.4. Automatic discovery of discriminative parts as a quadratic assignment problem

Participants: Ronan Sicre, Yannis Avrithis, Teddy Furon, Ewa Kijak.

Joint work with Julien Rabin and Frédéric Jurie (Univ. Caen).

Part-based image classification consists in representing categories by small sets of discriminative parts upon which a representation of the images is built. This piece of work addresses the question of how to automatically learn such parts from a set of labeled training images. We propose to cast the training of parts as a quadratic assignment problem in which optimal correspondences between image regions and parts are automatically learned. We analyze different assignment strategies and thoroughly evaluates them on two public datasets: Willow actions and MIT 67 scenes [45].

7.1.5. Learning DTW-preserving shapelets

Participants: Laurent Amsaleg, Arnaud Lods, Simon Malinowski.

Joint work with Romain Tavenard (Univ. Rennes 2).

Dynamic time warping (DTW) is one of the best similarity measures for time series, and it has extensively been used in retrieval, classification or mining applications. It is a costly measure, and applying it to numerous and/or very long times series is difficult in practice. Recently, shapelet transform (ST) proved to enable accurate supervised classification of time series. ST learns small subsequences that well discriminate classes, and transforms the time series into vectors lying in a metric space. We adopt the ST framework in a novel way: we focus on learning, without class label information, shapelets such that Euclidean distances in the ST-space approximate well the true DTW. Our approach leads to an ubiquitous representation of time series in a metric space, where any machine learning method (supervised or unsupervised) and indexing system can operate efficiently [28].

7.1.6. Tag propagation approaches within speaking face graphs for multimodal person discovery

Participants: Guillaume Gravier, Gabriel Sargent, Ronan Sicre.

Joint work with Gabriel Barbosa Da Fonseca, Izabela Lyon Freire, Zenilton Patrocinio Jr and Silvio Jamil F. Guimaraes (PUC Minas, Brazil)

The indexing of broadcast TV archives is a current problem in multimedia research. As the size of these databases grows continuously, meaningful features are needed to describe and connect their elements efficiently, such as the identification of speaking faces. In this context, we focused on two approaches for unsupervised person discovery. Initial tagging of speaking faces is provided by an OCR-based method, and these tags propagate through a graph model based on audiovisual relations between speaking faces. Two propagation methods are proposed, one based on random walks and the other based on a hierarchical approach. To better evaluate their performances, these methods were compared with two graph clustering baselines. We also study the impact of different modality fusions on the graph-based tag propagation scenario. From a quantitative analysis, we observed that the graph propagation techniques always outperform the baselines. Among all

compared strategies, the methods based on hierarchical propagation with late fusion and random walk with score-fusion obtained the highest MAP values. Finally, even though these two methods produce highly equivalent results according to Kappa coefficient, the random walk method performs better according to a paired t-test, and the computing time for the hierarchical propagation is more than 4 times lower than the one for the random walk propagation [22].

The tag propagation results were included in a large-scale comparison of systems for person discovery in broadcast videos resulting from the MediaEval 2016 international benchmark [27].

7.2. Multimedia content description and structuring

7.2.1. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality

Participant: Laurent Amsaleg.

Joint work with James Bailey, Dominique Barbe, Sarah Erfani, Michael Houle, Vinh Nguyen amd Miloš Radovanovic.

Recent research has shown that machine learning systems, including state-of-the-art deep neural networks, are vulnerable to adversarial attacks. By adding to the input object an imperceptible amount of adversarial noise, it is highly likely that the classifier can be tricked into assigning the modified object to any desired class. It has also been observed that these adversarial samples generalize well across models. A complete understanding of the nature of adversarial samples has not yet emerged. Towards this goal, we present a novel theoretical result formally linking the adversarial vulnerability of learning to the intrinsic dimensionality of the data. In particular, our investigation establishes that as the local intrinsic dimensionality (LID) increases, 1-NN classifiers become increasingly prone to being subverted. We show that in expectation, a k-nearest neighbor of a test point can be transformed into its 1-nearest neighbor by adding an amount of noise that diminishes as the LID increases. We also provide an experimental validation of the impact of LID on adversarial perturbation for both synthetic and real data, and discuss the implications of our result for general classifiers [13].

7.2.2. Efficient temporal kernels between feature sets for time series classification

Participant: Simon Malinowski.

Joint work with Romain Tavenard, Adeline Bailly, Louis Chapel (Univ. Rennes 2), Benjamin Bustos and Heider Sanchez (Univ. of Chile).

In the time-series classification context, the majority of the most accurate core methods are based on the bag-of-words framework, in which sets of local features are first extracted from time series. A dictionary of words is then learned and each time series is finally represented by a histogram of word occurrences. This representation induces a loss of information due to the quantization of features into words as all the time series are represented using the same fixed dictionary. In order to overcome this issue, we have designed a kernel operating directly on sets of features. Then, we have extended it to a time-compliant kernel that allows one to take into account the temporal information. We applied this kernel in the time series classification context. Proposed kernel has a quadratic complexity with the size of input feature sets, which is problematic when dealing with long time series. However, we have shown that kernel approximation techniques can be used to define a good trade-off between accuracy and complexity. We experimentally demonstrated that the proposed kernel can significantly improve the performance of time series classification algorithms based on bag-of-words [33].

7.2.3. Tampering detection and localization in images from social networks

Participants: Cédric Maigrot, Ewa Kijak, Vincent Claveau.

Verifying the authenticity of an image broadcast on social networks is crucial to limit the dissemination of false information. In this work, we aim to provide information about tampering localisation on such images, in order to help either the user or automatic methods to discriminate truth from falsehood. These images may have been subjected to a large number of possible forgeries, which calls for the use of generic methods. Image forensics methods based on local features have proven to be effective for the specific case of copy-move forgery. By taking advantage of the number of images available on the internet, we propose a generic system based on image retrieval, and image comparison based on local features to localise any kind of tampering in images from social networks.

Images from social media are likely to have undergone a large variety of modifications, some being malicious, and some not. The proposed approach is evaluated on three dedicated datasets containing a variety of representative tamperings in images from social media, with difficult examples. This allows an analysis of the local-features approaches behavior in this context. The method is further compared to several state-of-the-art methods and proves to be superior. Finally, we propose a classification step to discriminate malicious modifications from the non-malicious ones.

We have also built and made publicly available a large and challenging adapted database of real case images for evaluation [29].

7.2.4. Identity documents classification as an image classification problem

Participants: Ronan Sicre, Teddy Furon.

Joint work with Ahmad Montaser Awal and Nabil Ghanni (AriadNext).

This works studies the classification of images of identification documents. More specifically, we address the classification of documents composed of few textual information and complex background (such as identity documents). Unlike most existing systems, the proposed approach simultaneously locates the document and recognizes its class. The latter is defined by the document nature (passport, ID, etc.), emission country, version, and the visible side (main or back). This task is very challenging due to unconstrained capturing conditions, sparse textual information, and varying components that are irrelevant to the classification, e.g. photo, names, address, etc. First, a base of document models is created from reference images.

This problem is critical in various security context where proposed system must offer high performances. We address this challenge as an image classification problem, which has received a large attention from the scientific community. We show that training images are not necessary and only one reference image is enough to create a document model. Then, the query image is matched against all models in the base. Unknown documents are rejected using an estimated quality based on the extracted document. The matching process is optimized to guarantee an execution time independent from the number of document models. Once the document model is found, a more accurate matching is performed to locate the document and facilitate information extraction. Our system is evaluated on several datasets with up to 3042 real documents (representing 64 classes) achieving an accuracy of 96.6 % in [14].

In a second step, several methods are evaluated and we report results allowing a better understanding of the specificity of identification documents. We are especially interested in deep learning approaches, showing good transfer capabilities and high performances [44], [49].

7.2.5. Sentiment analysis

Participants: Vincent Claveau, Christian Raymond.

In the framework of the NexGenTV project, we have participated to the text-mining challenge DeFT about sentiment analysis. We have proposed methods for the identification of figurative language (irony, humor...), and for the classification of figurative and non-figurative tweets according to their polarity. For these tasks, we explore the use of three methods of increasing complexity: i) k-nearest neighbors with information retrieval based techniques, ii) boosting of decision trees, iii) recurrent neural networks [36]. It allows us to evaluate the precise interest of each of our approach and the data representation that they use: bag-of-words for the first one, n-grams for the second and word embedding for the latest.

7.3. Content-based information retrieval

7.3.1. Efficient diffusion on region manifolds: recovering small objects with compact CNN representations

Participants: Yannis Avrithis, Teddy Furon, Ahmet Iscen.

Joint work with Giorgos Tolias and Ondrej Chum (Technical University of Prague).

Query expansion is a popular method to improve the quality of image retrieval with both conventional and CNN representations. It has been so far limited to global image similarity. This work focuses on diffusion, a mechanism that captures the image manifold in the feature space. The diffusion is carried out on descriptors of overlapping image regions rather than on a global image descriptor like in previous approaches. An efficient off-line stage allows optional reduction in the number of stored regions. In the on-line stage, the proposed handling of unseen queries in the indexing stage removes additional computation to adjust the precomputed data. We perform diffusion through a sparse linear system solver, yielding practical query times well below one second. Experimentally, we observe a significant boost in performance of image retrieval with compact CNN descriptors on standard benchmarks, especially when the query object covers only a small part of the image. Small objects have been a common failure case of CNN-based retrieval [25].

7.3.2. Panorama to panorama matching for location recognition

Participants: Yannis Avrithis, Teddy Furon, Ahmet Iscen.

Joint work with Giorgos Tolias and Ondrej Chum (Technical University of Prague).

Location recognition is commonly treated as visual instance retrieval on "street view" imagery. The dataset items and queries are panoramic views, i.e., groups of images taken at a single location. This work introduces a novel panorama-to-panorama matching process, either by aggregating features of individual images in a group or by explicitly constructing a larger panorama. In either case, multiple views are used as queries. We reach near perfect location recognition on a standard benchmark with only four query views [26].

7.3.3. Memory vectors for similarity search in high-dimensional spaces

Participants: Teddy Furon, Ahmet Iscen.

Joint work with Vincent Gripon (IMT Atlantique), Michael Rabbat (Mc Gill University), and Hervé Jégou (Facebook AI Research).

We study an indexing architecture to store and search in a database of high-dimensional vectors from the perspective of statistical signal processing and decision theory. This architecture is composed of several memory units, each of which summarizes a fraction of the database by a single representative vector. The potential similarity of the query to one of the vectors stored in the memory unit is gauged by a simple correlation with the memory unit's representative vector. This representative optimizes the test of the following hypothesis: the query is independent from any vector in the memory unit vs. the query is a simple perturbation of one of the stored vectors. Compared to exhaustive search, our approach finds the most similar database vectors significantly faster without a noticeable reduction in search quality. Interestingly, the reduction of complexity is provably better in high-dimensional spaces. We empirically demonstrate its practical interest in a large-scale image search scenario with off-the-shelf state-of-the-art descriptors [6].

7.3.4. Exploiting multimodality in video hyperlinking to improve target diversity

Participants: Rémi Bois, Guillaume Gravier, Christian Raymond, Pascale Sébillot, Ronan Sicre, Vedran Vukotić.

Video hyperlinking is the process of creating links within a collection of videos to help navigation and information seeking. Starting from a given set of video segments, called anchors, a set of related segments, called targets, must be provided. In past years, a number of content-based approaches have been proposed with good results obtained by searching for target segments that are very similar to the anchor in terms of content and information. Unfortunately, relevance has been obtained to the expense of diversity. In this paper, we study multimodal approaches and their ability to provide a set of diverse yet relevant targets. We compare two recently introduced cross-modal approaches, namely, deep auto-encoders and bimodal LDA, and experimentally show that both provide significantly more diverse targets than a state-of-the-art baseline. Bimodal autoencoders offer the best trade-off between relevance and diversity, with bimodal LDA exhibiting slightly more diverse targets at a lower precision [17].

7.3.5. Generative adversarial networks for multimodal representation learning in video hyperlinking

Participants: Guillaume Gravier, Christian Raymond, Vedran Vukotić.

Continuous multimodal representations suitable for multimodal information retrieval are usually obtained with methods that heavily rely on multimodal autoencoders. In video hyperlinking, a task that aims at retrieving video segments, the state of the art is a variation of two interlocked networks working in opposing directions. These systems provide good multimodal embeddings and are also capable of translating from one representation space to the other. Operating on representation spaces, they lack the ability to operate in the original spaces (text or image), which makes it difficult to visualize the crossmodal function, and do not generalize well to unseen data. Recently, generative adversarial networks (GANs) have gained popularity and have been used for generating realistic synthetic data and for obtaining high-level, single-modal latent representation spaces. In this work, we evaluate the feasibility of using GANs to obtain multimodal representations. We show that GANs can be used for multimodal representation learning and that they provide multimodal representations that are superior to representations obtained with multimodal autoencoders. Additionally, we illustrate the ability of visualizing crossmodal translations that can provide human-interpretable insights on learned GAN-based video hyperlinking models [35].

7.4. Linking, navigation and analytics

7.4.1. Providing real-time insight during political debates in a second screen application

Participants: Vincent Claveau, Guillaume Gravier, Gabriel Sargent.

Joint work with Institut Eurecom, Wildmoka and AVISTO Telecom in the framework of the FUI project NexGenTV.

Second screen applications are becoming key for broadcasters exploiting the convergence of TV and Internet. Authoring such applications however remains costly. Within the NexGenTV project, we developed a second screen authoring application that leverages multimedia content analytics and social media monitoring. A back-office is dedicated to easy and fast content ingestion, segmentation, description and enrichment with links to entities and related content. From the back-end, broadcasters can push enriched content to front-end applications providing customers with highlights, entity and content links, overviews of social network, etc. The demonstration operates on political debates ingested during the 2017 French presidential election, enabling insights on the debates [12].

http://www.nexgentv.fr/communication/events/discover-our-new-politics-debates-live-video-edition

7.4.2. Information extraction in clinical documents

Participants: Clément Dalloux, Vincent Claveau.

Joint work with Claudia Moro (Pontifícia Universidade Católica do Paraná, Brazil) and Natalia Grabar (Univ. Lille)

Extracting fine-grained information from clinical texts is a keystone for numerous medical applications. For instance, in clinical trial protocols eligibility criteria are expressed through texts in an unstructured way. This year, we have developed an annotated corpus of clinical trials and made it available to the community. Based on this corpus, we proposed automatic methods to extract numerical information [20] and to handle the variation of the units used [43]. In such medical applications, detecting negation, uncertainty, and the scope on which they apply is important. Thus, we have also developed an annotated corpus, made it available to the community, and we have proposed an automatic tool based on recurrent neural networks [37], [41] and made it available as a web service.

7.4.3. Semi-supervision for information extraction

Participants: Vincent Claveau, Ewa Kijak.

Many NLP problems are tackled as supervised machine learning tasks. 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. This year, we examined and proposed such selection strategies in the specific case of conditional random fields (CRF) which are largely used in NLP. On the one hand, we proposed a simple method to correct a bias of some state-of-the-art selection techniques. On the other hand, we built an original approach to select the examples, based on the respect of proportions in the datasets. These contributions were validated over a large range of experiments implying several datasets and tasks, including named entity recognition, chunking, phonetization, word sense disambiguation [19].

7.4.4. Linking multimedia content for efficient news browsing via explorable news graphs

Participants: Rémi Bois, Guillaume Gravier, Pascale Sébillot.

Joint work with Maxime Robert, Éric Jamet (Univ. Rennes 2) and Emmanuel Morin (Univ. Nantes) in the framework of the CominLabs project Linking Media in Acceptable Hypergraphs.

As the amount of news information available online grows, media professionals are in need of advanced tools to explore the information surrounding specific events before writing their own piece of news, e.g., adding context and insight. While many tools exist to extract information from large datasets, they do not offer an easy way to gain insight from a news collection by browsing, going from article to article and viewing unaltered original content. Such browsing tools require the creation of rich underlying structures such as graph representations. These representations can be further enhanced by typing links that connect nodes, in order to inform the user on the nature of their relation. We propose an efficient way to generate links between news items in order to obtain an easily navigable graph, and enrich this graph by automatically typing created links. User evaluations are conducted on real-world data in order to assess for the interest of both the graph representation and link typing in a press reviewing task, showing a significant improvement compared to classical search engines [15], [16].

7.4.5. Multimodal detection of fake news

Participants: Vincent Claveau, Cédric Maigrot, Ewa Kijak.

Social networks make it possible to share rapidly and massively information, including fake news, hoaxes or rumors. Following our previous work in the frame of the Verification Multimedia Use task of Mediaeval 2016, we have explored the use of multimodal clues to detect fake news in social networks [38]. This year, we have studied the interest of combining and merging many approaches developed by the MediaEval participants in order to evaluate the predictive power of each modality. We have proposed several fusion strategies making the most of their potential complementarity [39].

7.5. Miscellaneous

In parallel with mainstream research activities, LINKMEDIA has a number of contribtions in other domains based on the expertise of the team members.

7.5.1. One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network

Participants: Guillaume Gravier, Christian Raymond, Vedran Vukotić.

Joint work with Silvia-Laura Pintea and Jan Van Gemert (TU Delft, The Netherlands).

There is an inherent need for autonomous cars, drones, and other robots to have a notion of how their environment behaves and to anticipate changes in the near future. In this work, we focus on anticipating future appearance given the current frame of a video. Existing work focuses on either predicting the future appearance as the next frame of a video, or predicting future motion as optical flow or motion trajectories starting from a single video frame. This work stretches the ability of convolutional neural networks (CNNs) to predict an anticipation of appearance at an arbitrarily given future time, not necessarily the next video frame. We condition our predicted future appearance on a continuous time variable that allows us to anticipate future frames at a given temporal distance, directly from the input video frame. We show that CNNs can learn an intrinsic representation of typical appearance changes over time and successfully generate realistic predictions at a deliberate time difference in the near future [34].

7.5.2. About zero bit watermarking error exponents

Participant: Teddy Furon.

This work aims to motivate more research works on the design of zero-bit watermarking schemes by showing an upper bound of the performances that known solutions failed to reach. To this end, an upper bound of error exponent characteristic is derived by translating Costa's rationale to zero-bit watermarking with side information. Three schemes are then considered: the dual-cone detection region originally proposed by Cox et al. and improved in Merhav et al. papers, ISS (Improved Spread Spectrum), and ZATT (Zero Attraction). It turns out that in certain conditions the latter performs better than the first one, which questions the optimality claimed Merhav et al. Nevertheless, the main conclusion is that these schemes are in general far away from the upper bound in the region of practical interest [23].

MIMETIC Project-Team

7. New Results

7.1. Outline

In 2017, MimeTIC has maintained his activity in motion analysis, modelling and simulation. In motion analysis, we focused our efforts on three major points: 1) being able to simplify the calibration and simulation of customized musculoskeletal models of the subjects, 2) explore how visual perception act on collision avoidance in pedestrian locomotion with an extension to group behavior, and 3) adapt accurate analysis in real condition (industrial or clinical contexts) where measurement inaccuracies and easy-to-use constraints make it difficult to directly apply methods used in laboratories.

From a long time, MimeTIC has been promoting the idea of using Virtual Reality to train human performance. On the one hand, it leads to an efficient tradeoff between high control and naturalness of the situation. On the other hand, it raises several fundamental questions about the automatic evaluation of the performance of the user, and the transfer of the skills trained in VR to real practice. In 2017, we explored these two questions by 1) developping new automatic methods for users' performance recognition and evaluation, and 2) biofidelity of mass manipulation in VR using haptic interfaces.

In virtual cinematography, we applied the analysis/synthesis approach to extract and simulate film styles and narration. We also extended our previously defiend Toric Space for camera placement to drone toric space to control a group of drones filming the action of an actor to ensure the coverture of cinematographic distinct viewpoints.

7.2. Motion analysis

7.2.1. Biomechanics for motion analysis-synthesis

Participants: Charles Pontonnier, Georges Dumont, Franck Multon, Antoine Muller, Diane Haering.

The PhD thesis of Antoine Muller defended on june, the 26 [1] aimed at democratizing the use of musculoskeletal analysis for a wide range of users. The work proposed contributions enabling better performances of such analyses and preserving accuracy, as well as contributions enabling an easy subject-specific model calibration. Firstly, in order to control the whole analysis process, the work is developed in a global approach of all the analysis steps: kinematics, dynamics [10] and muscle forces estimation. For all of these steps, quick analysis methods have been proposed. Particularly, a quick muscle force sharing problem resolution method [25] has been proposed, based on interpolated data. Moreover, a complete calibration process [24], based on classical motion analysis tools available in a biomechanical lab has been developed, based on motion capture and force platform data.

Diane Haering, Inria Post-doctoral fellow at MimeTIC works on the the determination of maximal torque enveloppes of the elbow. These results could have a great potential of application to quantify the articular load during work tasks [19], to help calibrating muscle parameters into musculoskeletal simulations [5]. The method has been integrated in a more global subject specific calibration method [30]. This method could also be used to better represent musculoskeletal models as in [6].

Ana-Lucia Cruz-Ruiz was a PhD student from november 2013 to december 2016. The goal of this thesis was to define and evaluate muscle-based controllers for motion control. This PhD was related to the ANR Entracte project. She developed an original control approach to reduce the redudancy of the musculoskeletal system. A low-dimensional representation of control mechanisms in throwing motions from a variety of subjects and target distances was proposed. The control representation stands at the kinematic level in task and joint spaces respectively, and at the muscle activation level using the theory of muscle synergies. Representative features were chosen and extracted using factorization and clustering techniques from the muscle data leading to better represent mechanisms hidden behind such dynamical motions, and could offer a promising control representation for synthesizing motions with muscle-driven characters [3].

7.2.2. Interactions between walkers

Participants: Anne-Hélène Olivier, Armel Crétual, Richard Kulpa, Sean Lynch, Laurentius Meerhoff.

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 to study both perception and action aspects of the interaction. Our efforts to validate the virtual reality platform to study interactions was acknowledged by a publication in IEEE TVCG 2017 [11] and was presented in IEEE VR 2017 conference [26]. Using the VR platform, we investigated the nature of visual information that is used for a collision free interaction. We aimed to manipulate the nature of visual information in two forms, global and local information appearances. The obstacle was presented with one of five virtual appearances, associated to global motion cues (i.e., a cylinder or a sphere), or local motion cues (i.e., only the legs or the trunk). A full body virtual walker, showing both local and global motion cues, used as a reference condition. The final crossing distance was affected by the global motion appearances, however, appearance had no qualitative effect on motion adaptations. These findings contribute towards further understanding what information people use when interacting with others. This work was published in TVCG 2017 [7] and presented as a poster in the ACAPS 2017 Conference[36]. This year, we also developed new experiments in our immersive platform. We designed a study to investigate the effect of gaze interception during collision avoidance between two walkers. In such a situation, mutual gaze can be considered as a form of nonverbal communication. Additionally, gaze is believed to detail future path intentions and to be part of the nonverbal negotiation to achieve avoidance collaboratively. We considered an avoidance task between a real subject and a virtual human character and studied the influence of the character's gaze direction on the avoidance behaviour of the participant. Virtual reality provided us with an accurate control of the situation: seventeen participants were immersed in a virtual environment, instructed to navigate across a virtual space using a joystick and to avoid a virtual character that would appear from either side. The character would either gaze or not towards the participant. Further, the character would either perform or not a reciprocal adaptation of its trajectory to avoid a potential collision with the participant. The findings of this paper were that during an orthogonal collision avoidance task, gaze behaviour did not influence the collision avoidance behaviour of the participants. Further, the addition of reciprocal collision avoidance with gaze did not modify the collision behaviour of participants. These results suggest that for the duration of interaction in such a task, body motion cues were sufficient for coordination and regulation. We discuss the possible exploitation of these results to improve the design of virtual characters for populated virtual environments and to interact with users. These results were presented to the AFRV 2017 conference [33] and submitted to IEEE VR 2018 conference.

We also provide lot of efforts to investigate, in collaboration with Julien Pettré from Inria Lagadic team, the process involved in the selection of interactions within our neighbourood. Considering the complex case of multiple interactions, we performed experiments in real conditions where a participant walked across a room whilst either one (i.e., pairwise) or two (i.e., group) participants crossed the room perpendicularly. By comparing these pairwise and group interactions, we assessed whether a participant avoids two upcoming collisions simultaneously, or as sequential pairwise interactions. Furthermore, in the group trials we varied the relative position of the two participants that crossed the trajectory of the other. This allowed us to change the affordance of passing through or around (i.e., its 'pass-ability'). Results showed that in the group trials, participants consistently avoided collision with lower risks of impending collision (as quantified by the future distance of closest approach) in the group compared to the pairwise trials. This implies that a participant – to some extent – interacted simultaneously with two other participants. Furthermore, we analysed in the group trials how the 'pass-ability' evolved over time. Results indicated that the affordance of passing through or around was already established early in the interaction. This shows that participants are susceptible to the affordance of passing through a gap between others. We concluded that pedestrians are able to interact with two other walkers simultaneously, rather than treating each interaction in sequence. These results were presented at the ICPA 2017 conference [21].

Finally, we continue working on the interaction between a walker and a moving robot. This work 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 just published in Gait and Posture our results on 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 was 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 [17]. Humans preferentially give way to the robot, even if this choice is not optimal with regard to motion adaptation to avoid the collision. In this new study, we considered the situation where the robot was reactive to the walker's motion. First of all, it results that humans have a good understanding of the robot behavior and their reaction are smoother and faster with respect to the case with a non-collaborative robot. Second, humans adapt similarly to human-human study and the crossing order is respected in almost all cases. These results have strong similarities with the ones observed with two humans crossing each other.

7.2.3. New automatic methods to assess motion in industrial contexts based on Kinect

Participants: Franck Multon, Georges Dumont, Charles Pontonnier, Pierre Plantard, Antoine Muller.

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 challenge 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. In this section, we detail the ergonomics application of such an approach. Firstly, in ergonomics, we explored the use of low-cost motion capture systems (i.e., 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. After having evaluated the actual accuracy of the pose reconstruction method delivered by the Kinect, we have identified that occlusions were a very important problem to solve in order to obtain reliable ergonomic assessments in real cluttered environments. To this end, we developed an approach to deal with long occlusions that occur in real manufacturing conditions. This approach is based on a structured database of examples (named filtered pose graph) that enables real-time correction of Kinect skeleton data [14].

This method has been applied to a complete ergonomic process outputting RULA scores based on the reconstructed and corrected poses. We challenged this method with a reference motion capture system in laboratory conditions [15]. To this end we compared joint angles and RULA scores obtained with our system and a reference Vicon mocap system in various conditions (with and without occlusions). The results show a very good accordance between manually tuned RULA scores given by experts and those computed by the automatic system. These results demonstrate that it could be used in industrial context to support the ergonomists decision-making process.

This year we also extended this work to evaluate if corrected data enabled us to estimate reliable joint torques using inverse dynamics to provide new information to ergonomic assessment [12]. Indeed, joint torques and forces are relevant quantities to estimate the biomechanical constraints of working tasks in ergonomics. However, inverse dynamics requires accurate motion capture data, which are generally not available in real manufacturing plants. Markerless and calibrationless measurement systems based on depth cameras, such as the Microsoft Kinect, are promising means to measure 3D poses in real time, such as using our corrected Kinect approach. Thus, we evaluated the reliability of an inverse dynamics method based on this corrected skeleton data and its potential use to estimate joint torques and forces in such cluttered environments. To this end, we compared the calculated joint torques with those obtained with a reference inverse dynamics method based on an optoelectronic motion capture system. Results show that the Kinect skeleton data enabled the inverse dynamics process to deliver reliable joint torques in occlusion-free (r=0.99 for the left shoulder elevation) and occluded (r=0.91 for the left shoulder elevation) environments. However, differences remain

between joint torques estimations. Such reliable joint torques open appealing perspectives for the use of new fatigue or solicitation indexes based on internal efforts measured on site. The study demonstrates that corrected Kinect data could be used to estimate internal joint torques, using an adapted inverse dynamics method. The method could be applied on-site because it can handle some cases with occlusions. The resulting Kinect-based method is easy-to-use, real-time and could assist ergonomists in risk evaluation on site.

This work was partially funded by the Faurecia company through a Cifre convention.

7.2.4. Clinical gait assessment based on Kinect data

Participant: Franck Multon.

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 lead 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. 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 enable the clinician to have a rapid state of the quality of the gait. We therefore proposed a new method to assess 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 artificially 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, 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.

This method has been challenged with other classical approaches to assess gait asymmetry using either cheap Kinect data or Vicon data. We demonstrate the superiority of the approach when using Kinect data for which traditional approaches failed to accurately detect gait asymmetry. It has been validated on healthy subjects who were forced to walk with a 5cm sole placed below each foot alternatively. In 2017 [2], we compared the results obtained with the famous Constant Relative Phase (CRP) that aims at quantifying within-stride asymmetry index. CRP requires noise-free and accurate motion capture, which is difficult to obtain in clinical settings. As our index, the Longitudinal Asymmetry Index (ILong), is obtained using data from a low-cost depth camera (Kinect) (depth images averaged over several gait cycles), rather than derived joint positions or angles, we checked that it could deliver more reliable asymmetry information within gait, compared to CRP. Hence, this study aimed to evaluate (1) the validity of CRP computed with Kinect, (2) the validity and sensitivity of ILong for measuring gait asymmetry based solely on data provided by a depth camera, (3) the clinical applicability of a posteriorly mounted camera system to avoid occlusion caused by the standard front-fitted treadmill consoles and (4) the number of strides needed to reliably calculate ILong. the results show that CRP based on times derivatives of joint angles failed to detect gait asymmetry, when using Kinect data. However, our index, ILong, detected this disturbed gait reliably and could be computed from a posteriorly placed Kinect without loss of validity. A minimum of five strides was needed to achieve a correlation coefficient of 0.9 between standard MBS and low-cost depth camera based ILong. ILong provides a clinically pragmatic method for measuring gait asymmetry, with application for improved patient care through enhanced disease, screening, diagnosis and monitoring.

This work has been done in collaboration with the MsKLab from Imperial College London, to design new gait asymmetry indexes that could be used in daily clinical analysis.

7.2.5. Biomechanical analysis of tennis serve

Participants: Caroline Martin, Richard Kulpa, Benoit Bideau, Pierre Touzard.

Following the previous studies we made on tennis serve, we were able to evaluate the link between performance and risk of injuries. To go further, we made new experiments on top-level young French players (between 12 up to 18 years old) to quantify the motor technical errors made (kinematics) and the impact on the risk of injury (dynamics). This experiments are part of a collaboration with the FFT (French Tennis Federation). We recently validated that the Waiter's serve implies higher risk of injuries [28]. It is a movement that was know by the coaches as not productive and risky but it was never validated.

7.3. Virtual human simulation

7.3.1. Novel Distance Geometry based approaches for Human Motion Retargeting

Since September 2016, Antonio Mucherino has a half-time Inria detachment in the MimeTIC team, in order to collaborate on exploring distance geometry-based problems in representing and editing human motion.

In this context, an extension of a distance geometry approach to dynamical problems was proposed in [23], and we co-supervised Antonin Bernardin for his Master thesis, which focused on applying such extended approach for retargeting human motions. In character animation, it is often the case that motions created or captured on a specific morphology need to be reused on characters having a different morphology. However, specific relationships such as body contacts or spatial relationships between body parts are often lost during this process, and existing approaches typically try to determine automatically which body part relationships should be preserved in such animation. Instead, we proposed a novel frame-based approach to motion retargeting [18], [22] which relies on a normalized representation of all the body joints distances to encompass all the relationships existing in a given motion. In particular, we proposed to abstract postures by computing all the inter-joint distances of each animation frame and to represent them by Euclidean Distance Matrices (EDMs). Such EDMs present the benefits of capturing all the subtle relationships between body parts, while being adaptable through a normalization process to create a morphology independent distance-based representation. Finally, they can also be used to efficiently compute retargeted joint positions best satisfying newly imposed distances. We demonstrated that normalized EDMs can be efficiently applied to a different skeletal morphology by using a dynamical distance geometry approach, and presented results on a selection of motions and skeletal morphologies.

In parallel, in collaboration with national (LIX, École Polytechnique, Palaiseau) and international partners, we have been working for improving the performances of existing algorithms for distance geometry, independently from the considered application. In [4], we analyzed the main causes for the approach to fail to provide accurate solutions in cases where interval distances are provided (instead of unique distance values), and we proposed some possible strategies to detect such situations. In [27], we presented a linear optimization problem for a common pre-processing step in distance geometry: the one of identifying a special vertex order allowing to discretize the solution search space.

7.4. Human motion in VR

7.4.1. Motion recognition and classification

Participants: Franck Multon, Richard Kulpa, Yacine Boulahia.

Action recognition based on human skeleton structure represents nowadays a prospering research field. This is mainly due to the recent advances in terms of capture technologies and skeleton extraction algorithms. In this context, we observed that 3D skeleton-based actions share several properties with handwritten symbols since they both result from a human performance. We accordingly hypothesize that the action recognition problem can take advantage of trial and error approaches already carried out on handwritten patterns. Therefore, inspired by one of the most efficient and compact handwriting feature-set, we proposed a skeleton descriptor referred to as Handwriting-Inspired Features. First of all, joint trajectories are preprocessed in order to handle the variability among actor's morphologies. Then we extract the HIF3D features from the processed joint locations according to a time partitioning scheme so as to additionally encode the temporal information over the sequence. Finally, we used Support Vector Machine (SVM) for classification. Evaluations conducted on two challenging datasets, namely HDM05 and UTKinect, testify the soundness of our approach as the obtained results outperform the state-of-the-art algorithms that rely on skeleton data [32].

This work has been carried-out in collaboration with the IRISA Intuidoc team, with Yacine Boulahia who is a co-supervised PhD student with Eric Anquetil.

7.4.2. Automatic evaluation of sports gesture

Participants: Richard Kulpa, Marion Morel.

Automatically evaluating and quantifying the performance of a player is a complex task since the important motion features to analyze depend on the type of performed action. But above all, this complexity is due to the variability of morphologies and styles of both the experts who perform the reference motions and the novices. Only based on a database of experts' motions and no additional knowledge, we propose an innovative 2-level DTW (Dynamic Time Warping) approach to temporally and spatially align the motions and extract the imperfections of the novice's performance for each joints [9]. We applied our method on tennis serve and karate katas [8].

7.4.3. Biofidelity in VR

Participants: Hilt Simon, Charles Pontonnier, Georges Dumont.

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 challenge 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. In this section, we detail the ergonomics application of such an approach. Firstly, in ergonomics, we explored the use of low-cost motion capture systems (i.e., a Microsoft Kinecte of geometrical and mechanical characteristics of the haptic device. Uncertainties on friction coefficients within the model are tuned thanks to an experimental protocol enabling a subjective comparison between real and virtual manipulations of a low mass object. The compensation of friction on the first and second axes of the haptic interface showed significant improvement of both realism and perceived load [20].

7.5. Digital story telling

7.5.1. Analysis of Film Style

Participants: Marc Christie, Hui-Yin Wu, Christophe Lino, Quentin Galvane.

We have designed and made available an open database of annotated film clips together with an analysis of elements of film style related to how the shots are composed, how the transitions are performed between shots and how the shots are sequenced to compose a film unit [29]. The purpose is to initiate a shared repository pertaining to elements of film style which can be used by computer scientists and film analysts alike. Though both research communities rely strongly on the availability of such information to foster their findings, current databases are either limited to low-level features (such as shots lengths, color and luminance information), contain noisy data, or are not available to the communities. The data and analysis we provide open exciting perspectives as to how computational approaches can rely more thoroughly on information and knowledge extracted from existing movies, and also provide a better understanding of how elements of style are arranged to construct a consistent message.

7.5.2. Film Editing Patterns: Thinking like a Director

Participants: Marc Christie, Hui-Yin Wu.

We have introduced *Film Editing Patterns (FEP)*, a language to formalize film editing practices and stylistic choices found in movies. FEP constructs are constraints expressed over one or more shots from a movie sequence [34] that characterize changes in cinematographic visual properties such as shot size, region, angle of on-screen actors.

We have designed the elements of the FEP language, then introduced its usage in annotated film data, and described how it can support users in the creative design of film sequences in 3D. More specifically: (i) we proposed the design of a tool to craft edited filmic sequences from 3D animated scenes that uses FEPs to support the user in selecting camera framings and editing choices that follow certain best practices used in cinema; (ii) we conducted an evaluation of the application with professional and non-professional filmmakers. The evaluation suggested that users generally appreciate the idea of FEP, and that it can effectively help novice and medium experienced users in crafting film sequences with little training and satisfying results.

7.5.3. Directing Cinematographic Drones

Participant: Marc Christie.

We have designed a set of high-level tools for filming dynamic targets with quadrotor drones. To this end, we proposed a specific camera parameter space (the Drone Toric space) together with interactive on-screen viewpoint manipulators compatible with the physical constraints of a drone. We then designed a real-time path planning approach in dynamic environments which ensures both cinematographic properties in viewpoints along the path and ensures the feasibility of the path by a quadrotor drone. We finally have demonstrated how the Drone Toric Space can be combined with our path planning technique to coordinate positions and motions of multiple drones around dynamic targets to ensure the coverture of cinematographic distinct viewpoints. The proposed research prototypes have been evaluation by an experienced drone pilot and filmmaker, as well as by non-experts users. Not only does the tool demonstrate it's benefit in rehearsing complex camera moves for the film and documentary industries, but it demonstrates it's usability for everyday recording of aethetic camera motions.

PANAMA Project-Team

7. New Results

7.1. Sparse Representations, Inverse Problems, and Dimension Reduction

Sparsity, low-rank, dimension-reduction, inverse problem, sparse recovery, scalability, compressive sensing

The team has had a substantial activity ranging from theoretical results to algorithmic design and software contributions in the fields of sparse representations, inverse problems, and dimension reduction.

7.1.1. Algorithmic and Theoretical results on Computational Representation Learning

Participants: Rémi Gribonval, Nicolas Bellot, Cássio Fraga Dantas.

Main collaborations: Luc Le Magoarou (IRT b<>com, Rennes), Nicolas Tremblay (GIPSA-Lab, Grenoble), R. R. Lopes and M. N. Da Costa (DSPCom, Univ. Campinas, Brazil)

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 literature to learn a dictionary from a collection of training samples, classical dictionary learning is limited to small-scale problems. Inspired by usual fast transforms, we proposed a general dictionary structure (called FAμST for Flexible Approximate Multilayer Sparse Transforms) that allows cheaper manipulation, and an algorithm to learn such dictionaries together with their fast implementation.

The principle and its application to image denoising appeared at ICASSP 2015 [80] and an application to speedup linear inverse problems was published at EUSIPCO 2015 [79]. A Matlab library has been released (see FAμST in Section 6.2) to reproduce the experiments from the comprehensive journal paper published in 2016 [82], which additionally includes theoretical results on the improved sample complexity of learning such dictionaries. Pioneering identifiability results have been obtained in the Ph.D. thesis of Luc Le Magoarou on this topic [83].

We further explored the application of this technique to obtain fast approximations of Graph Fourier Transforms. A conference paper on this latter topic appeared in ICASSP 2016 [81], and a journal paper has been published this year [17] where we empirically show that $\mathcal{O}(n \log n)$ approximate implementations of Graph Fourier Transforms are possible for certain families of graphs. This opens the way to substantial accelerations for Fourier Transforms on large graphs. The approximation error of such Fast Graph Fourier Transforms has been studied in a conference paper [31].

A C++ version of the FAμST software library has been developed (see Section 6) to release the resulting algorithms and interface them with both Matlab and Python (work in progress).

As a complement to the FAμST structure for matrix approximation, we proposed a learning algorithm that constrains the dictionary to be a sum of Kronecker products of smaller sub-dictionaries. A special case of the proposed structure is the widespread separable dictionary. This approach, named SuKro, was evaluated experimentally on an image denoising application [39].

We combined accelerated matrix-vector multiplications offered by FAuST matrix approximations with dynamic screening [52], that safely eliminates inactive variables to speedup iterative sparse recovery algorithms. First, we showed how to obtain safe screening rules for the exact problem while manipulating an approximate dictionary. We then adapted an existing screening rule to this new framework and define a general procedure to leverage the advantages of both strategies. Significant complexity reductions were obtained in comparison to screening rules alone [35].

7.1.2. Theoretical results on generalized matrix inverses, and the sparse pseudo-inverse

Participant: Rémi Gribonval.

Main collaboration: Ivan Dokmanic (University of Illinois at Urbana Champaign, USA)

We studied linear generalized inverses that minimize matrix norms. Such generalized inverses are famously represented by the Moore-Penrose pseudoinverse (MPP) which happens to minimize the Frobenius norm. Freeing up the degrees of freedom associated with Frobenius optimality enables us to promote other interesting properties. In a first part of this work [37], we looked at the basic properties of norm-minimizing generalized inverses, especially in terms of uniqueness and relation to the MPP. We first showed that the MPP minimizes many norms beyond those unitarily invariant, thus further bolstering its role as a robust choice in many situations. We then concentrated on some norms which are generally not minimized by the MPP, but whose minimization is relevant for linear inverse problems and sparse representations. In particular, we looked at mixed norms and the induced $\ell^p \to \ell^q$ norms.

An interesting representative is the sparse pseudoinverse which we studied in much more detail in a second part of this work [38], motivated by the idea to replace the Moore-Penrose pseudoinverse by a sparser generalized inverse which is in some sense well-behaved. Sparsity implies that it is faster to apply the resulting matrix; well-behavedness would imply that we do not lose much in stability with respect to the least-squares performance of the MPP. We first addressed questions of uniqueness and non-zero count of (putative) sparse pseudoinverses. We showed that a sparse pseudoinverse is generically unique, and that it indeed reaches optimal sparsity for almost all matrices. We then turned to proving a stability result: finite-size concentration bounds for the Frobenius norm of p-minimal inverses for $1 \leq p \leq 2$. Our proof is based on tools from convex analysis and random matrix theory, in particular the recently developed convex Gaussian min-max theorem. Along the way we proved several results about sparse representations and convex programming that were known folklore, but of which we could find no proof.

7.1.3. Algorithmic exploration of large-scale Compressive Learning via Sketching

Participants: Rémi Gribonval, Nicolas Keriven, Antoine Chatalic, Antoine Deleforge.

Main collaborations: Patrick Perez (Technicolor R&I France, Rennes), Anthony Bourrier (formerly Technicolor R&I France, Rennes; then GIPSA-Lab, Grenoble), Antoine Liutkus (ZENITH Inria project-team, Montpellier), Nicolas Tremblay (GIPSA-Lab, Grenoble), Phil Schniter & Evan Byrne (Ohio State University, USA)

Sketching for Large-Scale Mixture Estimation. When fitting a probability model to voluminous data, memory and computational time can become prohibitive. We proposed during the Ph.D. thesis of Anthony Bourrier [53], [56], [54], [55] 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 showed that it is possible to precisely estimate the mixture parameters provided that the sketch is large enough. The proposed algorithm provided good reconstruction and scaled to higher dimensions than previous probability mixture estimation algorithms, while consuming less memory in the case of voluminous datasets. It also provided a potentially privacy-preserving data analysis tool, since the sketch does not explicitly disclose information about individual datum.

During the Ph.D. thesis of Nicolas Keriven [12], we consolidated our extensions to non-isotropic Gaussians, with a new algorithm called CL-OMP [72] and conducted large-scale experiments demonstrating its potential for speaker verification. A conference paper appeared at ICASSP 2016 [71] and the journal version has been accepted this year [44], accompanied by a toolbox for reproducible research (see SketchMLBox, Section 6.3). Nicolas Keriven was awarded the SPARS 2017 Best Student Paper Award for this work [].

Sketching for Compressive Clustering and beyond. Last year we started a new endeavor to extend the approach beyond the case of Gaussian Mixture Estimation.

First, we showed empirically that sketching can be adapted to compress a training collection while still allowing large-scale *clustering*. The approach, called "Compressive K-means", uses CL-OMP at the learning stage and is described in a paper published at ICASSP 2017 [23]. In the high-dimensional setting, it is also possible to substantially speedup both the sketching stage and the learning stage by replacing Gaussian random matrices with fast random matrices in the sketching procedure. This has been demonstrated by Antoine

Chatalic during his internship and submitted for publication to a conference. An alternative algorithm for cluster recovery from a sketch was proposed this year, based on simplified hybrid generalized approximate message passing (SHyGAMP). Numerical experiments suggest that this approach is more efficient than CL-OMP (in both computational and sample complexity) and more efficient than k-means++ in certain regimes [25].

Then, we leveraged the parallel between the mathematical expression of sketched clustering and super-resolution to explore the potential of sketching and CL-OMP for the stable recovery of signals made of few spikes (in the gridless setting) from few random weighted Fourier measurements [47]. We also demonstrated that sketching can be used in blind source localization and separation, by learning mixtures of alpha-stable distributions [45], see details in Section 7.4.2 .

7.1.4. Theoretical results on Low-dimensional Representations, Inverse problems, and Dimension Reduction

Participants: Rémi Gribonval, Yann Traonmilin.

Main collaboration: Mike Davies (University of Edinburgh, UK), Gilles Puy (Technicolor R&I France, Rennes),

Inverse problems and compressive sensing 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). Last year we published a comprehensive paper [96] establishing generic RIP-based guarantees for the stable recovery of cones (positively homogeneous model sets) with arbitrary regularizers. We also described a generic technique to construct linear maps from a Hilbert space to $\mathbb{R}^m$ that satisfy the RIP [20]. These results have been surveyed in a book chapter completed this year[48].

Information preservation guarantees with low-dimensional sketches. We established a theoretical framework for sketched learning, encompassing statistical learning guarantees as well as dimension reduction guarantees. The framework provides theoretical grounds supporting the experimental success of our algorithmic approaches to compressive K-means, compressive Gaussian Mixture Modeling, as well as compressive Principal Component Analysis (PCA). A comprehensive preprint has been completed and submitted to a journal [42]. Future work will include expliciting the impact of the proposed framework on a wider set of concrete learning problems.

7.1.5. Algorithmic Exploration of Sparse Representations in Virtual Reality and Neurofeedback

Participant: Rémi Gribonval.

Ferran Argelaget & Anatole Lecuyer (HYBRID Inria project-team, Rennes), Saman Noorzadeh, Pierre Maurel & Christian Barillot (VISAGES Inria project-team, Rennes)

In collaboration with the VISAGES team we validated a technique to estimate brain neuronal activity by combining EEG and fMRI modalities in a joint framework exploiting sparsity [34].

Our work in collaboration with the HYBRID team on sparse dictionary learning for spatial and rotation invariant gesture recognition has been published this year [24]. Our work on multi-modal

7.1.5.1. An Alternative Framework for Sparse Representations: Sparse "Analysis" Models

Participants: Rémi Gribonval, Nancy Bertin, Clément Gaultier.

Main collaborations: Srdan Kitic (Technicolor R & I France, Rennes), Laurent Albera and Siouar Bensaid (LTSI, Univ. Rennes)

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 [70], [89][8], successful applications of this approach to sound source localization, brain imaging and audio restoration have been developed in the team during the last years [73], [75], [74], [50]. Along this line, two main achievements were obtained this year. First, and following the publication in 2016 of a journal paper embedding in a unified fashion our results in source localization [5], we wrote a book chapter (currently in press) gathering our contributions in physics-driven cosparse regularization, including new results and algorithms demonstrating the versatility, robustness and computational efficiency of our methods in realistic, large scale scenarios in acoustics and EEG signal processing [46]. Second, we continued extending the cosparse framework on audio restoration problems: improvements on our released real-time declipping algorithm (A-SPADE - see Section 6), new results on the denoising task [41], [28], and the submission of a journal paper encompassing several denoising and declipping methods in a common framework [40].

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), Adaptive Wavelet Packet Modulation (AWPM)

7.2.1. Characterizing and designing multi-carrier waveform systems with optimum PAPR

Participant: Rémi Gribonval.

Main collaboration: Marwa Chafii, Jacques Palicot, Carlos Bader (SCEE team, CentraleSupelec, 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 [57], defended in October 2016 and 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 [61] and several conference communications [59], [60]. Our characterization of waveforms with optimum PAPR [62] has been published in a journal in 2016 [58]. Our work on the design of new adaptive multi-carrier waveform systems able to cope with frequency-selective channels while minimizing PAPR which gave rise to a patent in 2016 [63] has been been submitted for publication as a journal paper. Our study of the tradeoffs between PAPR and Power Spectral Density properties of a wavelet modulation scheme has been published this year [14].

7.3. Emerging activities on Nonlinear Inverse Problems

Compressive sensing, compressive learning, audio inpainting, phase estimation

7.3.1. Locally-Linear Inverse Regression

Participant: Antoine Deleforge.

Main collaborations: Florence Forbes (MISTIS Inria project-team, Grenoble), Emeline Perthame (HUB team, Institut Pasteur, Paris), Vincent Drouard, Radu Horaud, Sileye Ba and Georgios Evangelidis (PERCEPTION Inria project-team, Grenoble)

A general problem in machine learning and statistics is that of *high- to low-dimensional mapping*. In other words, given two spaces $\mathbb{R}^D$ and $\mathbb{R}^L$ with $D \gg L$, how to find a relation between these two spaces such that given a new observation vector $y \in \mathbb{R}^D$ its associated vector $x \in \mathbb{R}^L$ can be estimated? In *regression*, a set of training pairs $\{(y_n, x_n)\}_{n=1}^N$ is used to learn the relation. In *dimensionality reduction*, only vectors $\{y_n\}_{n=1}^N$ are observed, and an intrinsic low-dimensional representation $\{x_n\}_{n=1}^N$ is sought. In [66], we introduced a probabilistic framework unifying both tasks referred to as *Gaussian Locally Linear Mapping* (GLLiM). The key idea is to learn an easier other-way-around locally-linear relationship from x to y using a joint Gaussian Mixture model on x and y. This mapping is then easily reversed via Bayes' inversion. This framework was notably applied to hyperspectral imaging of Mars [64], head pose estimation in images [16], sound source separation and localization [65], and virtually-supervised acoustic space learning (see Section 7.4.4). This year, in [19], we introduced the *Student Locally Linear Mapping* (SLLiM) framework. The use of heavy-tailed Student's t-distributions instead of Gaussian ones leads to more robustness and better regression performance on several datasets.

7.3.2. Phase Estimation in Multichannel Mixtures

Participants: Antoine Deleforge, Yann Traonmilin.

Main collaboration: Angélique Drémeau (ENSTA Bretagne and Lab-STICC, Brest)

The problem of estimating source signals given an observed multichannel mixture is fundamentally ill-posed when the mixing matrix is unknown or when the number of sources is larger that the number of microphones. Hence, prior information on the desired source signals must be incorporated in order to tackle it. An important line of research in audio source separation over the past decade consists in using a model of the source signals' magnitudes in the short-time Fourier domain [9]. Such models can be inferred through, *e.g.*, non-negative matrix factorization [9] or deep neural networks [90]. Magnitudes estimates are often interpreted as instantaneous variances of Gaussian-process source signals, and are combined with Wiener filtering for source separation. In [26], we introduced a shift of this paradigm by considering the *Phase Unmixing* problem: how can one recover the instantaneous phases of complex mixed source signals when their magnitudes and mixing matrix are known? This problem was showed to be NP-hard, and three approaches were proposed to tackle it: a heuristic method, an alternate minimization method, and a convex relaxation into a semi-definite program. The last two approaches were showed to outperform the oracle multichannel Wiener filter in under-determined informed source separation tasks. The latter yielded best results, including the potential for exact source separation in under-determined settings. In [27] we applied this framework to the classical problem of *phase retrieval* with a novel multivariate Von Mises prior on phases. We showed that enforcing this prior yielded more accurate estimates than state-of-the art phase retrieval methods.

7.3.3. Audio Inpainting and Denoising

Participants: Rémi Gribonval, Nancy Bertin, Clément Gaultier.

Main collaborations: Srdan Kitic (Technicolor R&I France, Rennes)

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 [49]), we proposed over the last two years a series of algorithms leveraging the competitive cosparse approach, which offers a very appealing trade-off between reconstruction performance and computational time [74], [77] [6]. The work on cosparse audio declipping which was awarded the Conexant best paper award at the LVA/ICA 2015 conference [77] resulted in a software release in 2016.

In 2017, this work was extended towards advanced (co)sparse decompositions, including several forms of structured sparsity in the time-frequency domain and across channels, and towards their application to the denoising task, in addition to the previously introduced declipping task, which we continued to improve. In particular, we investigated the incorporation of the so-called "social" structure constraint [78] into problems regularized by a cosparse prior [28], [41], and exhibited a common framework allowing to tackle both

denoising and declipping in a unified fashion [40]. A new algorithm for joint declipping of multichannel audio was also derived (one submitted conference publication.)

7.4. Source Localization and Separation

Source separation, sparse representations, probabilistic model, source localization

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, which is the task of retrieving the source signals underlying a multichannel mixture signal.

Over the last years, we proposed a general probabilistic framework for the joint exploitation of spatial and spectral cues [9], hereafter summarized as the "local Gaussian modeling", and 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. This model became the basis of a large number of works in the field, including our own. This accumulated progress lead, in 2015, to two main achievements: a new version of the Flexible Audio Source Separation Toolbox, fully reimplemented, was released [93] and we published an overview paper on recent and going research along the path of *guided* separation in a special issue of IEEE Signal Processing Magazine [10].

From there, our recent work divided into several tracks: maturity work on the concrete use of these tools and principles in real-world scenarios, in particular within the voiceHome and INVATE projects (see Section 7.4.1) ; more exploratory work towards new approaches diverging away from local Gaussian modeling (Section 7.4.2) ; formulating and addressing a larger class of problems related to localization and separation, in the context of robotics (Section 7.4.3) and audio scene analysis with machine learning (Section 7.4.4).

7.4.1. Towards Real-world Separation and Remixing Applications

Participants: Nancy Bertin, Frédéric Bimbot, Rémi Gribonval, Ewen Camberlein, Romain Lebarbenchon, Mohammed Hafsati.

Main collaborations: Emmanuel Vincent (MULTISPEECH Inria project-team, Nancy), Nicolas Epain (IRT b<>com, Rennes)

Based on the team's accumulated expertise and tools for localization and separation using the local Gaussian model, two real-world applications were addressed in the past year, which in turn gave rise to new research tracks.

First, we were part of the voiceHome project (2015-2017, see Section 9.1.4), an industrial collaboration aiming at developing natural language dialog in home applications, such as control of domotic and multimedia devices, in realistic and challenging situations (very noisy and reverberant environments, distant microphones). We benchmarked, improved and optimized existing localization and separation tools to the particular context of this application, worked on a better interface between source localization and source separations steps and on optimal initialization scenarios, and reduced the latency and computational burden of the previously available tools, highlighting operating conditions were real-time processing is achievable. Automatic selection of the best microphones subset in an array was investigated. A journal publication including new data (extending the voiceHome Corpus, see Section 6.1), baseline tools and results was submitted to a special issue of Speech Communication. Accomplished progress and levers of improvements identified thanks to this project resulted in the granting of an Inria ADT (Action de Développement Technologique), which started in September 2017, for a new development phase of the FASST software (see Section 6.5).

Second, through the Ph.D. of Mohammed Hafsati (in collaboration with the IRT b<>com with the INVATE project, see Section 9.1.2) started in November 2016, we investigated a new application of source separation to sound re-spatialization from Higher Order Ambisonics (HOA) signals [69], in the context of free navigation in 3D audiovisual contents. We studied the applicability conditions of the FASST framework to HOA signals and benchmarked localization and separation methods in this domain. We started extending our methods to hybrid acquisition scenarios, where the separation of HOA signals can be informed by the complementary close-up microphonic signals. Future work will include systematic experimental evaluation.

7.4.2. Beyond the Local Complex Gaussian Model

Participants: Antoine Deleforge, Nicolas Keriven.

Main collaboration: Antoine Liutkus (ZENITH Inria project-team, Montpellier)

The team has also recently investigated a number of alternative probabilistic models to the local complex Gaussian (LCG) model for audio source separation. An important limit of LCG is that most signals of interest such as speech or music do not exhibit Gaussian distributions but heavier-tailed ones due to their important dynamic [84]. In [45] we proposed a new sound source separation algorithm using heavy-tailed alpha stable priors for source signals. Experiments showed that it outperformed baseline Gaussian-based methods on under-determined speech or music mixtures. Another limitation of LCG is that it implies a zero-mean complex prior on source signals. This induces a bias towards low signal energies, in particular in under-determined settings. With the development of accurate magnitude spectrogram models for audio signals such as nonnegative matrix factorization [91][9] or more recently deep neural networks [90], it becomes desirable to use probabilistic models enforcing strong magnitude priors. In [26], we explored deterministic magnitude models (see section 7.3.2 for details). An approximate and tractable probabilistic version of this referred to as BEADS (Bayesian Expansion Approximating the Donut Shape) is currently under development. The source prior considered is a mixture of isotropic Gaussians regularly placed on a zero-centered complex circle.

7.4.3. Applications to Robot Audition

Participants: Nancy Bertin, Antoine Deleforge, Martin Strauss, Victor Miguet.

Main collaborations: Aly Magassouba, Pol Mordel and François Chaumette (LAGADIC Inria project-team, Rennes), Alexander Schmidt and Walter Kellermann (University of Erlangen-Nuremberg, Germany)

Implicit Localization through Audio-based Control. 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 investigated in the context of Aly Magassouba's Ph.D. (defended in December 2016) focused on a sensor-based control approach. A journal paper encompassing and extending the results obtained before 2017 [88], [86], [87] has been submitted to IEEE Transactions on Robotics (accepted with minor revisions). In 2017, we obtained new results on the use of interaural level difference as the only input feature of the servo, with new experimental validation on humanoid robots. A publication about these last results has been submitted to IEEE Robotics and Automation Letters.

Ego-noise Reduction with Motor-Data-Guided Dictionary Learning. Ego-noise reduction is the problem of suppressing the noise a robot caused by its own motions. Such noise degrades the recorded microphone signal such that the robot's auditory capabilities suffer. To suppress it, it is intuitive to use also motor data, since it provides additional information about the robot's joints and thereby the noise sources. In [95], we incorporated motor data to a recently proposed multichannel dictionary algorithm [68]. We applied this to ego-noise reduction on the humanoid robot NAO. At training, a dictionary is learned that captures spatial and spectral characteristics of ego-noise. At testing, nonlinear classifiers are used to efficiently associate the current robot's motor state to relevant sets of entries in the learned dictionary. By this, computational load is reduced by one third in typical scenarios while achieving at least the same noise reduction performance. Moreover, we proposed to train dictionaries on different microphone array geometries and used them for ego-noise reduction while the head on which the microphones are mounted is moving. In such scenarios, the motor-data-guided approach resulted in significantly better performance values.

Sound Source Localization with a Drone. Flying robots or drones have undergone a massive development in recent years. Already broadly commercialized for entertainment purpose, they also underpin a number of exciting future applications such as mail delivery, smart agriculture, archaeology or search and rescue. An important technological challenge for these platforms is that of localizing sound sources in order to better analyse and understand their environment. For instance, how to localize a person crying for help in the context

of a natural disaster? This challenge raises a number of difficult scientific questions. How to efficiently embed a microphone array on a drone? How to deal with the heavy ego-noise produced by the drone's motors? How to deal with moving microphones and distant sources? Victor Miguet and Martin Strauss tackled part of these challenges during their masters' internships. A light 3D-printed structure was designed to embed a USB sound card and a cubic 8-microphone array under a Mikrokopter drone that can carry up to 800 g of payload in flights. Noiseless speech and on-flights ego-noise datasets were recorded. The data were precisely annotated with the target source's position, the state of each drone's propellers and the drone's position and velocity. Baseline methods including multichannel Wiener filtering, GCC-PHAT and MUSIC were implemented in both C++ and Matlab and were tested on the dataset. Up to 5° speech localization accuracy in both azimuth and elevation was achieved under heavy-noise conditions (-5 dB signal-to-noise-ratio). We plan to make the datasets and code publicly available in 2018.

7.4.4. Virtually-Supervised Auditory Scene Analysis

Participants: Antoine Deleforge, Nancy Bertin, Diego Di Carlo, Clément Gaultier.

Main collaborations: Ivan Dokmanic (University of Illinois at Urbana-Champaign, Coordinated Science Lab, USA) and Robin Scheibler (Tokyo Metropolitan University, Tokyo, Japan), Saurabh Kataria (IIT Kanpur, India)

Classical audio signal processing methods strongly rely on a good knowledge of the *geometry* of the audio scene, *i.e.*, what are the positions of the sources, the sensors, and how does the sound propagate between them. The most commonly used *free field* geometrical model assumes that the microphone configuration is perfectly known and that the sound propagates as a single plane wave from each source to each sensor (no reflection or interference). This model is not valid in realistic scenarios where the environment may be unknown, cluttered, dynamic, and include multiple sources, diffuse sounds, noise and/or reverberations. Such difficulties critical hinders sound source separation and localization tasks. In some ongoing work, we showed that the knowledge of a few early acoustic echoes significantly improve sound source separation performance over the free-field model.

Recently, two directions for advanced audio geometry estimation have emerged and were investigated in our team. The first one is physics-driven [46]. This approach explicitly solves the wave propagation equation in a given simplified yet realistic environment assuming that only few sound sources are present, in order to recover the positions of sources, sensors, or even some of the wall absorption properties. Encouraging results were obtained in simulated settings, including "hearing behind walls" [76]. However, these methods rely on approximate models and on partial knowledge of the system (e.g. room dimensions), limiting their real-world applicability so far. The second direction is data-driven. It uses machine learning to bypass the use of a physical model by directly estimating a mapping from acoustic features to source positions, using training data obtained in a real room [65], [67]. These methods can in principle work in arbitrarily complex environments, but they require carefully annotated training datasets. Since obtaining such data is time consuming, the methods are usually working well for one specific room and setup, and are hard to generalize in practice.

We proposed a new paradigm that aims at making the best of physics-driven and data-driven approaches, referred to as *virtually acoustic space travelling* (VAST) [22], [30]. The idea is to use a physics-based room-acoustic simulator to generate arbitrary large datasets of room-impulse responses corresponding to various acoustic environments, adapted to the physical audio system at hand. We demonstrated that mappings learned from these data could potentially be used to not only estimate the 3D position of a source but also some acoustical properties of the room [30]. We also showed that a virtually-learned mapping could robustly localize sound sources from real-world binaural input, which is the first result of this kind in audio source localization [22]. The starting PhD thesis of Diego Di Carlo aims at applying the VAST framework to the blind estimation of acoustic echoes. The ultimate goal is to use these estimates to recover partial acoustic properties of the scene and enhance audio signal processing methods.

7.5. Music Content Processing and Information Retrieval

Music structure, music language modeling, System & Contrast model, complexity

Current work developed in our research group in the domain of music content processing and information retrieval explore various information-theoretic frameworks for music structure analysis and description [51], in particular the System & Contrast model [1].

7.5.1. Tensor-based Representation of Sectional Units in Music

Participants: Corentin Guichaoua, Frédéric Bimbot.

Following Kolmogorov's complexity paradigm, modeling the structure of a musical segment can be addressed by searching for the compression program that describes as economically as possible the musical content of that segment, within a given family of compression schemes.

In this general framework, packing the musical data in a tensor-derived representation enables to decompose the structure into two components : (i) the shape of the tensor which characterizes the way in which the musical elements are arranged in an n-dimensional space and (ii) the values within the tensor which reflect the content of the musical segment and minimize the complexity of the relations between its elements.

This approach has been studied in the context of Corentin Guichaoua's PhD [11] where a novel method for the inference of musical structure based on the optimisation of a tensorial compression criterion has been designed and experimented.

This tensorial compression criterion exploits the redundancy resulting from repetitions, similarities, progressions and analogies within musical segments in order to pack musical information observed at different timescales in a single n-dimensional object.

The proposed method has been introduced from a formal point of view and has been related to the System & Constrast Model [1] as a extension of that model to hypercubic tensorial patterns and their deformations.

From the experimental point of view, the method has been tested on 100 pop music pieces (RWC Pop database) represented as chord sequences, with the goal to locate the boundaries of structural segments on the basis of chord grouping by minimizing the complexity criterion. The results have clearly established the relevance of the tensorial compression approach, with F-measure scores reaching 70 %

7.5.2. Modeling music by polytopic graphs of latent relations

Participants: Corentin Louboutin, Frédéric Bimbot.

The musical content observed at a given instant within a music segment obviously tends to share privileged relationships with its immediate past, hence the sequential perception of the music flow. But local music content also relates with distant events which have occurred in the longer term past, especially at instants which are metrically homologous (in previous bars, motifs, phrases, etc.) This is particularly evident in strongly "patterned" music, such as pop music, where recurrence and regularity play a central role in the design of cyclic musical repetitions, anticipations and surprises.

The web of musical elements can be described as a Polytopic Graph of Latent Relations (PGLR) which models relationships developing predominantly between homologous elements within the metrical grid.

For regular segments the PGLR lives on an n-dimensional cube(square, cube, tesseract, etc...), n being the number of scales considered simultaneously in the multiscale model. By extension, the PGLR can be generalized to a more or less regular n-dimensional polytopes.

Each vertex in the polytope corresponds to a low-scale musical element, each edge represents a relationship between two vertices and each face forms an elementary system of relationships.

The estimation of the PGLR structure of a musical segment can be obtained computationally as the joint estimation of the description of the polytope, the nesting configuration of the graph over the polytope (reflecting the flow of dependencies and interactions between the elements within the musical segment) and the set of relations between the nodes of the graph, with potentially multiple possibilities.

If musical elements are chords, relations can be inferred by minimal transport [85] defined as the shortest displacement of notes, in semitones, between a pair of chords. Other chord representations and relations are possible, as studied in [33] where the PGLR approach is presented conceptually and algorithmically, together with an extensive evaluation on a large set of chord sequences from the RWC Pop corpus (100 pop songs).

Specific graph configurations, called Primer Preserving Permutations (PPP) are extensively studied in [32] and are related to 6 main redundant sequences which can be viewed as canonical multiscale structural patterns.

These results illustrate the efficiency of the proposed model in capturing structural information within musical data and is currently being explored on melodic sequences and rythmic patterns.

7.5.3. Regularity Constraints for the Fusion of Music Structure Segmentation System

Participant: Frédéric Bimbot.

Main collaborations Gabriel Sargent (LinkMedia Inria project-team, Rennes)

Music structure estimation has become a central topic within the field of Music Information Retrieval. Indeed, as music is a highly structured information stream, knowledge of how a music piece is organized represents a key challenge to enhance the management and exploitation of large music collections.

Former work carried out in our group [94] has illustrated the benefits that can be expected from a regularity constraint on the structural segmentation of popular music pieces : a constraint which favors structural segments of comparable size provides a better conditioning of the boundary estimation process.

As a further investigation, we have explored the benefits of the regularity constraint as an efficient way for combining the outputs of a selection of systems presented at MIREX between 2010 and 2015. These experiments have yielded a level of performance which is competitive to that of the state-of-the-art on the ”MIREX10" dataset (100 J-Pop songs from the RWC database) [21].

SIROCCO Project-Team

7. New Results

7.1. Analysis and modeling for compact representation

3D modelling, light-fields, 3D meshes, epitomes, image-based rendering, inpainting, view synthesis

7.1.1. Visual attention

Participant: 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 deal with the understanding and modeling of overt attention.

Saccadic model: Since 2015, we have worked on saccadic model, which predicts the visual scanpaths of an observer watching a scene displayed onscreen. In 2016, we proposed a first improvement consisting in using spatially-variant and context-dependent viewing biases. We showed that the joint distribution of saccade amplitudes and orientations is significantly dependent on the type of visual stimulus. In addition, the joint distribution turns out to be spatially variant within the scene frame. This model outperforms state-of-the-art saliency models, and provides scanpaths in close agreement with human behavior. In [19], [35], we went further by showing that saccadic models are a flexible framework that can be tailored to emulate observer's viewing tendencies. More specifically, we tailored the proposed model to simulate visual scanpaths of 5 age groups of observers (i.e. adults, 8-10 y.o., 6-8 y.o., 4-6 y.o. and 2 y.o.). The key point is that the joint distribution of saccade amplitude and orientation is a visual signature specific to each age group, and can be used to generate age-dependent scanpaths. Our age-dependent saccadic model does not only output human-like, age-specific visual scanpaths, but also significantly outperforms other state-of-the-art saliency models. We demonstrated that the computational modelling of visual attention, through the use of saccadic model, can be efficiently adapted to emulate the gaze behavior of a specific group of observers.

Effects on Comics by Clustering Gaze Data: Comics are a compelling communication medium conveying a visual storytelling. With a smart mixture of text or/and other visual information, artists tell a story by drawing the viewer attention on specific areas. With the digital comics revolution (e.g. mobile comic and webcomic), we are witnessed a resurgence of interest for this art form. This new form of comics allows not only to tackle a wider audience but also new consumption methods. An open question in this endeavor is identifying where in a comic panel the effects should be placed. We proposed a fast, semi-automatic technique to identify effects-worthy segments in a comic panel by utilizing gaze locations as a proxy for the importance of a region. We took advantage of the fact that comic artists influence viewer gaze towards narrative important regions. By capturing gaze locations from multiple viewers, we can identify important regions. The key contribution is to leverage a theoretical breakthrough in the computer networks community towards robust and meaningful clustering of gaze locations into semantic regions, without needing the user to specify the number of clusters. We have developed a method based on the concept of relative eigen quality that takes a scanned comic image and a set of gaze points and produces an image segmentation. A variety of effects such as defocus, recoloring, stereoscopy, and animations has been demonstrated. We also investigated the use of artificially generated gaze locations from saliency models in place of actual gaze locations.

Perceptual metric for perceptual transfer: Color transfer between input and target images has raised a lot of interest in the past decade. 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. Different color transfer methods often result in different output images. The process of determining the most plausible output image is difficult and requires, due to the lack of an objective metric, time-consuming and costly subjective experiments.

To overcome this problem, we proposed a perceptual model for evaluating results from color transfer methods [31]. From a subjective experiment, involving several color transfer methods, we build a regression model with random forests to describe the relationship between a set of features (e.g. objective quality, saliency, etc.) and the subjective scores. An analysis and a cross-validation showed that the predictions of the proposed quality metric are highly accurate.

7.1.2. Saliency-based navigation in omnidirectional image

Participants: Olivier Le Meur, Thomas Maugey.

Omnidirectional images describe the color information at a given position from all directions. Affordable 360° cameras have recently been developed leading to an explosion of the 360 degrees data shared on the social networks. However, an omnidirectional image does not contain interesting content everywhere. Some part of the images are indeed more likely to be looked at by some users than others. Knowing these regions of interest might be useful for 360° image compression, streaming, retargeting or even editing. In the work published in [25], a new approach based on 2D image saliency is proposed both to model the user navigation within a 360° image, and to detect which parts of an omnidirectional content might draw users' attention. A double cube projection is first used to put the saliency estimation in the classical 2D image framework. Consecutively, the saliency map serves as a support for the navigation estimation algorithm.

Figure 2. Example of the saliency map (left) and the estimated navigation (right).

7.1.3. Context-aware Clustering and Assessment of Photo Collections

Participants: Dmitry Kuzovkin, Olivier Le Meur.

To ensure that all important moments of an event are represented and that challenging scenes are correctly captured, both amateur and professional photographers often opt for taking large quantities of photographs. As such, they are faced with the tedious task of organizing large collections and selecting the best images among similar variants. Automatic methods assisting with this task are based on independent assessment approaches, evaluating each image apart from other images in the collection. However, the overall quality of photo collections can largely vary due to user skills and other factors. We explore the possibility of context-aware image quality assessment, where the photo context is defined using a clustering approach, and statistics of both the extracted context and the entire photo collection are used to guide identification of low-quality photos. We demonstrate that the proposed method is able to adapt flexibly to the nature of processed albums and to facilitate the task of image selection in diverse scenarios.

7.1.4. Light fields view extraction from lenslet images

Participants: Pierre David, Christine Guillemot, Mikael Le Pendu.

Practical systems have recently emerged for the capture of real light fields which go from cameras arrays to single cameras mounted on moving gantries and plenoptic cameras. While camera arrays capture the scene from different viewpoints, hence with a large baseline, plenoptic cameras use an array of micro-lenses placed in front of the photosensor to separate the light rays striking each microlens into a small image on the photosensors pixels, and this way capture dense angular information with a small baseline. Extracting views

from the raw lenslet data captured by plenoptic cameras involves several processing steps: devignetting which, with white images, aims at compensating for the loss of illumination at the periphery of the micro-lenses, color demosaicing, alignment of the sensor data with the micro-lens array, and converting the hexagonal sampling grid into a rectangular sampling grid. These steps are quite critical as they have a strong impact on the quality of the extracted sub-aperture images (views).

We have addressed two important steps of the view extraction from lenslet data: color demosaicing and alignment of the micro-lens array on the photosensor. We have developed a new method guided by a white lenslet image for color demosaicing of raw lenslet data [27](best paper award). The white lenslet image gives measures of confidence on the color values which are then used to weight the color samples interpolation (see Fig.3 . Similarly, the white image is used to guide the interpolation performed in the alignment of the micro-len arrays on the photosensor. The method significantly decreases the crosstalk artefacts from which suffer existing methods.

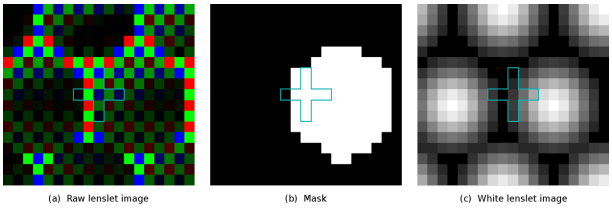

Figure 3. (a) is the raw image we want to demosaic, (b) is a mask which holds every pixel belonging to the same lenslet, (c) white image.

7.1.5. Super-rays for efficient Light fields processing

Participants: Matthieu Hog, Christine Guillemot.

Light field acquisition devices allow capturing scenes with unmatched post-processing possibilities. However, the huge amount of high dimensional data poses challenging problems to light field processing in interactive time. In order to enable light field processing with a tractable complexity, we have addressed, in collaboration with Neus Sabater (technicolor) the problem of light field over-segmentation [15]. We have introduced the concept of super-ray, which is a grouping of rays within and across views (see Fig.4), as a key component of a light field processing pipeline. The proposed approach is simple, fast, accurate, easily parallelisable, and does not need a dense depth estimation. We have demonstrated experimentally the efficiency of the proposed approach on real and synthetic datasets, for sparsely and densely sampled light fields. As super-rays capture a coarse scene geometry information, we have also shown how they can be used for real time light field segmentation and correcting refocusing angular aliasing.

7.2. Representation and compression of large volumes of visual data

Sparse representations, data dimensionality reduction, compression, scalability, perceptual coding, rate-distortion theory

7.2.1. Cloud-based image and video compression

Participants: Jean Begaint, Christine Guillemot.

Figure 4. Super-rays for the sparsely sampled light field in the Tsukuba dataset.

The emergence of cloud applications and web services has led to an increasing use of online resources for storing and exchanging images and videos. Billions of images are already stored in the cloud, and hundreds of millions are uploaded every day. Redundancy between images stored in the cloud can be leveraged to efficiently compress images by exploiting inter-images correlations. We have developed a region-based prediction scheme to exploit correlation between images in the cloud. In order to compensate the deformations between correlated images, the reference image of the cloud is first segmented into multiple regions determined from matched local features and aggregated super-pixels. We then estimate a photometric and geometric deformation model between the matched regions in the reference frame and frame to be coded. Multiple references are then generated, by applying the estimated deformation models to the reference frame, and organized in a pseudo-sequence to be differentially encoded with classic video coding tools. Experimental results demonstrate that the proposed approach yields significant rate-distortion performance improvements compared to current coding solutions such as HEVC.

7.2.2. 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. The goal of the collaboration with Ericsson is to develop novel compression algorithms that allow 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. In 2016, a scalable video compression was developed 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. In 2017 instead, we developed a backward compatible compression algorithm of HDR images, where only the LDR data are sent [14]. The novelty of the approach relies on the optimization of an invertible mapping called Tone Mapping Operator (TMO) that maps efficiently the HDR data to the LDR data. Two optimizations have been carried out in a rate-distortion sense: in the first problem, the distortion of the HDR data is minimized under the constraint of minimum LDR datarate, while in the second problem, a new constraint is added in the optimization problem to insure that LDR data are closed to some "aesthetic" a priori. Taking into account the aesthetic of the scene in video compression is indeed novel, since video compression is traditionally optimized to deliver the smallest distortion with the input data at the minimum datarate. Moreover, we provided new statistical models for estimating the distortions and the rate and showed

their accuracy to the real data. Finally, a novel axis is currently carried out to efficiently exploit the temporal redundancy in HDR videos.

7.2.3. Sparse image representation and deep learning for compression

Participants: Thierry Dumas, Christine Guillemot, Aline Roumy.

Deep learning is a novel research area that attempts to extract high level abstractions from data by using a graph with multiple layers. One could therefore expect that deep learning might allow efficient image compression based on these high level features. However, there are many issues that make the learning task difficult in the context of image compression. First, learning a transform is equivalent to learning an autoencoder, which is of its essence unsupervised and therefore more difficult that classical supervised learning, where deep learning has shown tremendous results. Second, the learning has to be performed under a rate-distortion criterion, and not only a distortion criterion, as is classically done in machine learning. Last but not least, deep learning, as classical machine learning, consists in two phases: (i) build a graph that can make a good representation of the data (i.e. find an architecture usually made with neural nets), and (ii) learn the parameters of this architecture from large-scale data. As a consequence, neural nets are well suited for a specific task (text or image recognition) and require one training per task. The difficulty to apply machine learning approach to image compression is that it is important to deal with a large variety of patches, and with also various compression rates. Different architectures have been proposed to design a single neural network that can work efficiently at any coding rate either by a Winner Take all approach [28] or an adaptation to the quantization noise during the training [40].

7.2.4. 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.

In 2016, we have developed a graph-based representation for complex camera configurations. In particular, a generalized Graph-Based Representation has beend eveloped which handles two views with complex translations and rotations between them. The proposed approach uses the epipolar segments to have a row-wise description of the geometry that is as simple as for rectified views. In 2017, the Graph-based Representation has been extended to build a rate-distortion optimized description of the geometry of multi-view images [22]. This work brings two major novelties. First the graph can now handle multiple views (more than 2) thanks to a recursive construction of the geometry across the views. Second, the number of edges describing the geometry information is carefully chosen with respect to a rate-distortion criterion evaluated on the reconstructed views.

An adaptation of the graph-based representations (GBR) has been proposed to describe color and geometry information of light fields (LF) in [38]. Graph connections describing scene geometry capture inter-view dependencies. They are used as the support of a weighted Graph Fourier Transform (wGFT) to encode disoccluded pixels. The quality of the LF reconstructed from the graph is enhanced by adding extra color

information to the representation for a sub-set of sub-aperture images. Experiments show that the proposed scheme yields rate-distortion gains compared with HEVC based compression (directly compressing the LF as a video sequence by HEVC).

7.2.5. Light fields compression using sparse reconstruction

Participants: Fatma Hawary, Christine Guillemot.

Light field data exhibits large amount of information, which poses challenging problems in terms of storage capacity, hence the need for efficient compression schemes. In collaboration with Technicolor (Dominique Thoreau and Guillaume Boisson), we have developed a scalable coding method for the light field data based on the sparsity of light fields in the angular (view) domain. A selected set of the light field sub-aperture images is encoded as a video sequence in a base layer and transmitted to the decoder. The remaining light field views are then reconstructed from the decoded subset of views, by exploiting the light field sparsity in the angular continuous Fourier domain. The reconstructed light field is enhanced using a patch-based restoration method which further exploits the light field angular redundancy.

7.2.6. Light fields dimensionality reduction and compression

Participants: Elian Dib, Christine Guillemot, Xiaoran Jiang, Mikael Le Pendu.

We have investigated low rank approximation methods exploiting data geometry for dimensionality reduction of light fields. We have developed an approximation method in which homographies and the rank approximation model are jointly optimized [32]. The homographies are searched in order to align linearly correlated sub-aperture images in such a way that the batch of views can be approximated by a low rank model. The light field views are aligned using either one global homography or multiple homographies depending on how much the disparity across views varies from one depth plane to the other. The rank constraint is expressed as a product of two matrices, where one matrix contains basis vectors and where the other one contains weighting coefficients. The basis vectors and weighting coefficients can be compressed separately exploiting their respective characteristics. The optimization hence proceeds by iteratively searching for the homographies and the factored model of the input set of sub-aperture images (views), which will minimize the approximation error.

A light field compression algorithm based on a low rank approximation exploiting scene and data geometry has then be developed [18]. The best pair of key parameters (approximation rank and quantization step size), in terms of rate-distortion performance, of the algorithm are predicted based on a model learned from a set of training light fields. The model is learned as a function of several input light field features: disparity indicators defined as a function of the decay rate of the SVD values of the original and registered view matrices, as well as texture indicators defined in terms of the decay rate of SVD values computed on the central view. The parameter prediction problem is cast as a multi-output classification problem solved using a Decision Tree ensemble method, namely the Random Forest method. The approximation method is currently being extended to local super-ray based low rank models.

7.3. Rendering, inpainting and super-resolution

image-based rendering, inpainting, view synthesis, super-resolution

7.3.1. Transformation of the Beta distribution for color transfer

Participants: Hristina Hristova, Olivier Le Meur.

After having investigated the use of multivariate generalized Gaussian distribution in color transfer, we propose a novel transformation between two Beta distributions. The key point is that performing a Gaussian-based transformation between bounded distributions may result in out-of-range values. Furthermore, as a symmetrical distribution, the Gaussian distribution cannot model asymmetric distributions. This reveals important limitations of the Gaussian model when applied to image processing tasks and, in particular, to color transfer. To tackle these limitations of the Gaussian-based transformations, we investigate the use of bounded distributions, and more specifically, the Beta distribution. The Beta distribution is a bounded two-parameter

dependent distribution, which can admit different shapes and thus, fit various data, bounded in a discrete interval. Adopting the Beta distribution to model color and light distributions of images is our key idea and motivation. The proposed transformation progressively and accurately reshapes an input Beta distribution into a target Beta distribution using four intermediate statistical transformations. Experiments have shown that the proposed method obtains more natural and less saturated results than results of recent state-of-the-art color transfer methods. Moreover, the results portray better both the target color palette and the target contrast.

7.3.2. Light field inpainting and edit propagation

Participants: Oriel Frigo, Christine Guillemot, Mikael Le Pendu.

With the increasing popularity of computational photography brought by light field, simple and intuitive editing of light field images is becoming a feature of high interest for users. Light field editing can be combined with the traditional refocusing feature, allowing a user to include or remove objects from the scene, change its color, its contrast or other features.

A simple approach for editing a light field image can be obtained with an edit propagation, where first a particular subaperture view is edited (most likely the center one) and then a coherent propagation of this edit is performed through the other views. This problem is particularly challenging for the task of inpainting, as the disparity field is unknown under the occludding mask.

We have developed two methods which exploit two different light field priors, namely a low rank prior and a smoothness prior in epipolar plane images (EPI) to propagate a central view inpainting or edit to all the other views. In the first method, a set of warped versions of the inpainted central view with random homographies are vectorized and concatenated columnwise into a matrix together with the views of the light field to be inpainted. Because of the redundancy between the views, the matrix satisfies a low rank assumption enabling us to fill the region to inpaint with low rank matrix completion. To this end, a new matrix completion algorithm, better suited to the inpainting application than existing methods, has also been developed. In its simple form, our method does not require any depth prior, unlike most existing light field inpainting algorithms. The method has then been extended to better handle the case where the area to inpaint contains depth discontinuities.

In the second approach, the problem of propagating an edit from a single view to the remaining light field is solved by a structure tensor driven diffusion on the epipolar plane images [29]. Since EPIs are piecewise smooth and have no complex texture content, tensor driven diffusion is naturally suited for inpainting the EPIs as an efficient technique to obtain a coherent edit propagation. The proposed method has been shown to be useful for two applications: light field inpainting and recolorization. While the light field recolorization is obtained with a straightforward diffusion, the inpainting application is particularly challenging, as the structure tensors accounting for disparities are unknown under the occluding mask. This issue has been addressed with a disparity inpainting by means of an interpolation constrained by superpixel boundaries.

7.3.3. Light fields super-resolution

Participants: Christine Guillemot, Lara Younes.

Capturing high spatial resolution light fields remains technologically challenging, and the images rendered from real light fields have today a significantly lower spatial resolution compared to traditional 2D cameras. In collaboration with the University of Malta (Prof. Reuben Farrugia), we have developed an example-based super-resolution algorithm for light fields, which allows the increase of the spatial resolution of the different views in a consistent manner across all sub-aperture images of the light field [12]. To maintain consistency across all sub-aperture images of the light field, the algorithm operates on 3D stacks (called patch-volumes) of 2D-patches, extracted from the different sub-aperture images. The patches forming the 3D stack are best matches across subaperture images. A dictionary of examples is first constructed by extracting, from a training set of high- and low- resolution light fields, pairs of high- and low-resolution patch-volumes. These patch-volumes are of very high dimension. Nevertheless, they contain a lot of redundant information, hence actually lie on subspaces of lower dimension. The low- and high-resolution patch-volumes of each pair can therefore be projected on their respective low and high-resolution subspaces using e.g. Principal Component Analysis (PCA). The dictionary of pairs of projected patch-volumes (the examples) map locally the relation between the

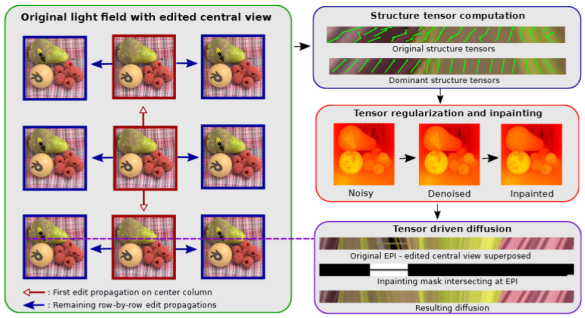

Figure 5. Overview of the proposed method. On the left, a light field with an inpainted central view and after a first edit propagation, which is performed through the center column of views (red arrows). Remaining edit propagations are performed row-by-row (blue arrows). The remaining boxes illustrate the steps of epipolar plane diffusion: Structure tensor computation where we obtain dominant structure tensors which are then spatially regularized and inpainted to estimate the structre tensors in the unknown part of the light field. Finally a tensor driven diffusion is performed on the EPIs.

high-resolution patch volumes and their low-resolution (LR) counterparts. A linear mapping function is then learned, using Multivariate Ridge Regression (RR), between the subspaces of the low- and high- resolution patch-volumes. Each overlapping patch-volume of the low-resolution light field can then be super-resolved by a straight application of the learned mapping function (some results in Fig.6).

Figure 6. Illustration of some results on a crop of a central view with a magnification factor of 2; Top row: original (left), bicubic interpolation (right); Bottom row: super-resolved with a deep learning technique of the litterature (left), super-resolved with the proposed method (right).

This work is currently being extended on one hand by exploring how deep learning techniques can further benefit the scheme, and on the other hand by considering a hybrid system in which a 2D high resolution image, a priori not aligned with the light field views, can guide the light field super-resolution process.

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. Interactive Coding for Navigation in 3D scenes (ICON 3D)

Participants: Thomas Maugey, Aline Roumy.

In order to have performing FTV systems, the data transmission has to take into account the interactivity of the user, *i.e.,* the viewpoint that is requested. In other words, a FTV system transmits to the visualisation support only what needs to be updated when a user changes its viewpoint angle (*i.e.*, the new information appearing in its vision field).

In the context of the project ICON 3D funded by the GdR-Isis, we have developed new geometry prediction algorithms for surface meshes. Given a part of a mesh, the prediction algorithm is able to estimate a neighboring mesh subset corresponding to the one newly visible after user viewpoint angle change. For each mesh of a 3D model, we have generated all the predictions possible depending on the part of the model known by the decoder. Then we have characterized the prediction error.

The question of which data representation to use for Interactive Navigation has also been studied in [20]. More precisely, the navigation domain is split in small segments, each of them coded independently. This work has developed some optimal partitioning solution for different navigation scenario.

7.4.2. Correlation model selection for interactive video communication

Participants: Navid Mahmoudian Bidgoli, Thomas Maugey, Aline Roumy.

Interactive video communication has been recently proposed for multi-view videos. In this scheme, the server has to store the views as compactly as possible while allowing interactive navigation. Interactive navigation refers to the possibility for the user to select one view or a subset of views. To achieve this goal, the compression must be done using a model-based coding in which the correlation between the predicted view generated on the user side and the original view has to be modeled by a statistical distribution. In the context of the project Intercomm, the work published in [37] has proposed a framework for lossless fixed-length source coding to select a model among a candidate set of models that incurs the lowest extra rate cost to the system. Moreover, in cases where the depth image is available, we provide a method to estimate the correlation model.

7.4.3. Optimal selection of reference sensors for spatially correlated data storage

Participants: Thomas Maugey, Aline Roumy.

Highly instrumented Smart-cities, which are now common urban policies, are facing problems of management and storage of a large volume of data coming from an increasing number of sources. In the context of the project Intercom, we have proposed a data compression method by predictive coding of spatially correlated multi-source data. In a nutshell, some sensors are selected as references. They are used to predict the other sensor values, based on a Kriging prediction. We have proposed an algorithm to optimally select both the number and the position of the reference sensors among all the ones that are stored on a server and shared with a high number of users. This work has been done in collaboration with the Inria I4S project-team, IFFSTAR and the L2S.